Design of Protein-Protein Interaction Specificity Using Computational Methods and Experimental Library Screening

by

Tsan-Chou Scott Chen

B.S. Chemistry
National Taiwan University, 2004

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2012

Signature of Author: _____
Department of Biology
February 03, 2012

Certified by: _____
Amy Keating
Associate Professor of Biology
Thesis Supervisor

Accepted by: _____
Robert T. Sauer
Salvador E. Luria Professor of Biology
Co-Chair, Biology Graduate Committee

Design of Protein-Protein Interaction Specificity Using Computational Methods and Experimental Library Screening

by

Tsan-Chou Scott Chen

Submitted to the Department of Biology
on February 03, 2012 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biology at the Massachusetts Institute of Technology

ABSTRACT

Computational design of protein-protein interaction specificity is a powerful tool to examine and expand our understanding about how protein sequence determines interaction specificity. It also has many applications in basic bioscience and biotechnology. One of the major challenges for design is that current scoring functions relying on general physical principles do not always make reliable predictions about interaction specificity. In this thesis I described application of two approaches to address this problem. The first approach sought to improve scoring functions with experimental interaction specificity data related to the protein family of design interest. I used this approach to design inhibitor peptides against the viral bZIP protein BZLF1. Specificity against design self-interaction was considered in the study. The second approach exploited the power of experimental library screening to characterize a large number of designed sequences at once, increasing the overall probability of identifying successful designs. I presented a novel framework for such library design approach and applied it to the design of anti-apoptotic Bcl-2 proteins with novel interaction specificity toward BH3 peptides. Finally I proposed how these two approaches can be combined together to further enhance our design capabilities.

Thesis Supervisor: Amy Keating
Title: Associate Professor of Biology

For my parents

ACKNOWLEDGEMENTS

I would like to first thank my advisor Amy Keating for her mentoring through the past five years. In addition to her helpful guidance, she created a great working environment. In this environment, I enjoyed the freedom to pursue my research interest and to interact with other great colleagues. I would also like to thank my thesis committee members, Bob Sauer, Chris Burge and Bruce Tidor for helpful suggestions on my research projects as well as my career goals.

I have the privilege to have worked with many great people in the Kearting lab. Thanks to Orr Ashenberg, Judy Baek, Joe DeBartolo, Sanjib Dutta, Emiko Fire, Glenna Foight, Xiaoran Fu, Gevorg Grigoryan, Karl Gutwin, Seungsoo Hahn, Karl Hauschild, Jen Kaplan, Yong Ho Kim, Chris Negron, Hector Palacious, Vladimir Potapov, Luther Reich, Aaron Reinke, Emzo de los Santos, Michael Schneider, Josh Sims, Ben Steele, Evan Thompson and Nora Zizlsperger for their help and company over the past few years. I would like to thank Gevorg Grigoryan for helping me to get familiarize with the field of computational protein design. A significant portion of my research has been inspired by his work in the Keating lab. I want to thank Aaron Reinke for our collaborative effort on the BZLF1 inhibitor design project, and for our conversation about science in general. I want to thank Sanjib Dutta for teaching me almost every aspect of the yeast surface display method for my library design work. I want to thank Evan Thompson for really helpful input on the BZLF1 inhibitor design project as well as his often very useful advice on experimental troubleshooting. I want to thank Hector Palacious, who worked with me with great enthusiasm for a summer on the Bcl-2 library design project. I want to thank both Karl Gutwin and Vladimir Potapov for their help on computer issues. Finally, I want to thank Orr Ashenberg, Jen Kaplan, Chris Negron and Seungsoo Hahn for a great amount of fun time we shared together, in or out of the lab.

My graduate career has been made possible because of my family and friends. My parents and my sister Stephanie have been really supportive of all my pursuits. There is no substitute for their love and caring. And last but not least, I want to thank my wife Sidney, for simply everything.

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

Protein-protein interactions play important roles in virtually every aspect of cell biology. The functional significance of protein-protein interactions suggests that interactions have likely evolved to occur only between selected protein partners. The highly specific nature of protein-protein interaction has indeed been demonstrated *in vitro* and *in vivo*[1] in numerous studies. Recent efforts aiming to profile protein-protein interactions comprehensively in selected organisms have further provided a picture of such specificity on a global scale[2,3,4].

One key question to protein biochemists is how protein-protein interaction specificity is achieved. Assays studying protein-protein interactions in isolation of their native environment, especially those done using purified proteins, have clearly demonstrated that interaction specificity can be encoded within proteins themselves[5,6]. This further suggests that protein-protein interaction specificity can be encoded at the primary sequence level, as protein primary sequence determines three-dimensional structure[7], which in turns determines interaction properties in solution.

It is perhaps not surprising that proteins with very different sequences can fold into different structures and have distinct interaction properties. However, many recent studies have revealed that proteins (or protein domains/motifs) highly similar in sequence and/or structure can possess very different interaction specificities as well. These observations suggest that diverse interaction specificity can be evolved from a common protein sequence family/structural fold. A major implication is that interaction networks can be evolved with increasing complexity without the

need to reinvent all important protein components from scratch. Examples include modular domains such as the PDZ[5,8,9,10], Src homology 2 (SH2) [11,12] and the Src homology 3 (SH3) [13,14] families that are present in many cell signaling proteins, the coiled-coil motifs of different bZIP transcription factors[6], the Bcl-2 proteins involved in apoptosis[15], and cell-adhesion molecules such as the *Drosophila* protein Dscam[16]. Dscam represents a particularly interesting case from an evolutionary perspective. Dscam consists of 10 immunoglobulin-like domains, three of which are variable and play important roles in homodimerization. Each of these 3 variable domains is encoded by an exon block, and mutually exclusive splicing at each block gives rise to more than 10,000 distinct isoforms. It was shown that each variable domain is largely specific for interaction with itself, and their combined action results in high hemophilic binding specificity for the full-length Dscam[16], a key property for neurons to distinguish self from non-self (self-avoidance) in development. Evolutionary analysis suggested that each exon block was evolved by exon duplication followed by sequence divergence[17], illustrating how selective pressure exerted by the desire to maintain self-avoidance can help shape the remarkable homo-specificity for the Dscam family.

Interestingly, solved structures of proteins with similar sequences but distinct interaction specificities have revealed that often the same binding interface is utilized, and specificity can be attributed to local differences in structures[18,19,20,21,22,23,24,25,26] (Fig 1.1). One mechanism by which natural proteins can change their interaction properties but still preserve a common protein fold is to adopt different conformations for loops linking helices/strands that define the basic scaffold. A good example is the SH2 family, which interacts with peptides that contain a phosphorylated tyrosine (pTyr)[11,12]. Specificity of SH2 domains interacting with different pTyr peptides is crucial for specificity in transmitting signals from protein tyrosine kinases to their downstream

pathways. Three main specificity classes of SH2 domains have been discovered that recognize peptides with different sequence signatures C-terminal to pTyr, including ones with an Asn at the second residue C-terminal to pTyr (P+2), ones with a hydrophobic residue at P+3, and ones with a hydrophobic residue at P+4. Loops flanking the binding interface for SH2 domains confer selectivity toward these 3 types of peptides by opening or blocking binding pockets for the P+2, P+3 or P+4 residues[19,20]. The SH3 family also utilizes different loop conformations at the binding interface to provide specificity toward different peptides that are rich in prolines[19]. Antibodies provide another classic example of using loops to confer different binding properties[23], sharing a common immunoglobulin scaffold but using variation in 6 surface loops, the complementarity-determining regions (CDR), to achieve exquisite specificities for their target antigens.

Although larger changes in local conformations such as loops present a convenient way to change interaction properties, examples of more subtle sequence/structural features providing specificity abound in nature as well. One such example is the interaction between colicins endonuclease (DNases) and immunity (Im) proteins. Colicins are stress-induced *E.coli* bacteriocins that target other *E.coli* cells. Toxicity of colicins against their own producing cells can be neutralized by binding to their cognate Im proteins, so high interaction specificity is critical. A crystal structure of a non-cognate complex between E9 DNase and Im2 was solved in a recent study and comparison was made to the structure of the cognate complex between E9 DNase and Im9[24]. It was observed that backbone and sidechain packing at the core of the two interfaces was highly similar. However, the presence of unfavorable polar/charged residue burial and sub-optimal hydrogen bonding patterns weakened interaction significantly for the non-cognate complex. For bZIP coiled coils, structural and mutational analysis have revealed

presence of specificity features described by particular patterns of hydrophobic packing, hydrogen bonding, and electrostatic interactions between different sidechain pairs[25]. Interaction can be encoded through combination of these features, without any significant change in backbone structure. For the SH2 and SH3 domains discussed in the previous paragraph, it has also been shown that structural features similar to the ones described above are important to further fine-tune the specificity obtained from loops[12,14,19].

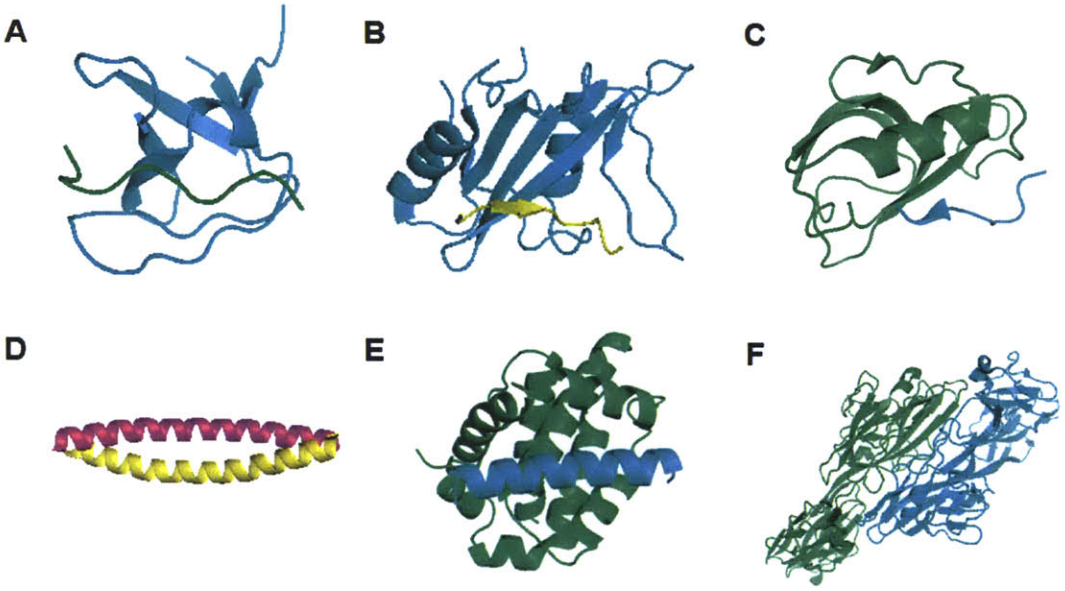


Figure 1-1 Examples of protein-protein interactions that achieve specificity using a common structural fold.
A representative complex structure was shown for each class of interactions: (A) Complex between a SH3 domain from the Abl tyrosine kinase and a proline-rich peptide. (PDB ID: 1ABO)[156]. (B) Complex between a SH2 domain from the SAP protein and a phosphotyrosine peptide (PDB ID: 1D4W)[157]. (C) Complex between an Erbin PDZ domain and the C-terminal tail of the ErbB2 receptor (PDB ID: 1MFG)[158]. (D) Complex between the bZIP coiled coil motifs of FOS and JUN (PDB ID: 1FOS)[159]. (E) Complex between the anti-apoptotic Bcl-2 protein Mcl-1 and the BH3-only peptide Bim (PDB ID: 2PQK)[160]. (F) Complex of a homodimer formed by the N-terminal domain of a particular Dscam isoform[161].

Examining different strategies used by nature to achieve interaction specificity offers the exciting possibility that one can mimic or devise new strategies to design interaction specificity. In fact, many attempts have been made to change interaction specificity for the different protein systems described above by altering loops or simply introducing one or a few amino acid mutations at their binding interfaces[27,28,29,30]. Given the importance of protein-protein interaction, the ability to design protein-protein interaction specificity could find many applications in the study of cell biology[31]. Proteins have been redesigned to create dominant negatives/potential therapeutics specific for the target[32,33,34], to generate obligate heterodimers[35,36], to test the functional significance of the many different interactions of an original protein[37,38], and to create novel interactions to rewire cell signaling in synthetic biology applications[39,40]. In addition to these applications, evaluating the success/failure of designs can help examine and advance our understanding of how protein primary sequence influences interaction specificity.

Traditionally, researchers have attempted design using rather general knowledge obtained from structural/mutational analysis of protein-protein interactions. For example, the importance of hydrogen bonding, favorable electrostatics, and shape complementarity are well known[41]. Nonetheless, this approach often fails to capture the complexity/subtlety of how sequence influences interaction specificity. Experimental alanine scanning[42,43] and hydrophile scanning[44] have also been used to generate proteins/peptides with novel interaction specificities, but the chemical diversity accessible by such approaches is rather limited.

Recent technologies, both computational and experimental, have helped revolutionize the field of protein design. Below I first introduce the concept of computational protein design and its application to designing protein-protein interaction specificity. I then survey the field of

experimental library screening, with particular focus on how its combination with computational protein design could become a powerful approach moving forward.

Computational protein design

The idea behind computational protein design stems from the principle that because protein primary sequence determines protein function in solution, one should be able to develop a quantitative understanding of the relation between sequence and function[45,46,47]. In the context of protein-protein interaction[48,49], function refers to the free energy change ($\Delta\Delta G$) of the protein interacting with its partner. If such a relation could be computed, one could perform computational instead of experimental searches through the vast sequence space in order to identify sequences with the desired properties.

Scoring function

Different types of scoring models have been developed to compute energy from sequence. Among them, physics based structural modeling is the most general, as it aims to address the question using basic physical principles[50]. In this approach, the structure of a protein/protein complex is first predicted from sequence, and an energy score is computed from the structure. The complexity of protein molecules largely prevents their description directly by quantum mechanics except in special instances[51,52], and the energy is usually evaluated as a combination of molecular mechanics terms such as van der Waals and Coulombic electrostatics[53]. Solvation of the protein is often approximated by a polar and a non-polar component to avoid the computationally expensive cost of treating water molecules explicitly[54]. The polar component addresses (1) the screening of electrostatic interactions within the protein by the surrounding solvent and (2) the energetic cost of burying and thus desolvating a charged or polar amino acid

side chain in a folded protein. The polar component is often computed using a continuum electrostatics model[55,56,57]. The non-polar component attempts to describe the hydrophobic effect and can be approximated using terms that depend on solvent accessible surface area, or other methods[58,59,60,61,62]. Not surprisingly, the necessity to include different approximations for physics based models affects their accuracies. This will be discussed in more detail later.

In contrast to physics-based models, statistical potentials aim to estimate energies using statistics derived from the PDB. These potentials usually approximate the energy as the sum of terms describing interactions between two residues[63] or two atoms[64,65] contacting each other in the structure being evaluated, although attempts to capture higher order interactions have been reported[66,67]. The score for a particular residue-residue or atom-atom interaction is based on the observed number of such contacts from the PDB, corrected for by the expected number of random encounters for the same residue or atom pair. A contact can be defined simply on a distance basis[64,65], but more sophisticated potentials take into account other information such as orientation [68,69] or the environment of the contact as well[70]. One advantage of statistical potentials is their speed compared to physics based structural models. In addition, it is known that current physics based structural models do not always accurately describe the geometry of interactions between different residues or atoms observed from the PDB. Examples include certain packing preferences among hydrophobic sidechains[71] and the angle distribution of hydrogen bonds[72]. On the other hand, statistical potentials have their own big approximations in converting observed statistics into energies[73], and by no means present a more physically meaningful formulation than physics based models. Statistical potential have been used in multiple prediction and design problems[74] including protein-protein interactions[75,76]. Potentials

that include terms from both physics based models and statistical potentials have also been develop (e.g. Rosetta[77]) and have had great success in many applications.

Design objectives

After choosing the scoring function, one must define the objective(s) for which a designed sequence would be optimized. One could optimize a protein-protein interaction by minimizing the interaction energy (ΔΔG), calculated by subtracting the energy of each protein partner modeled in its unbound forms from the energy of the modeled complex. Approximations such as a rigid–body docking or even less formal definitions are sometimes employed, due to difficulty in modeling the unbound reference states[78,79].

For designing protein-protein interaction specificity, one often needs to consider interactions with one or more undesired proteins, in addition to that with the target. Determining beforehand how specific the designed protein should be when predicted computationally (i.e. the predicted energy gap between binding the target and binding the undesired proteins) is not straightforward. As scoring functions are not always accurate, it is tempting to pick designed proteins predicted to be highly specific as a larger predicted specificity gap is likely more tolerant of errors in prediction. However, it is known that trade-offs can exist between affinity toward the target and specificity against the undesired proteins[80,81,82]. Focusing only on widening the specificity gap can therefore create designed proteins that are specific but bind the target weakly. Depending on the application, it can be beneficial to explore a range of different designs with different trade-offs in affinity and specificity. More specific examples of this will be presented later.

Search in structure and sequence space

Guided by a scoring function and the objectives, the next critical component in design is to search for sequences with desired properties among an immense and combinatorial space. Because evaluating a sequence requires the determination of its optimal structure, the search needs to be performed in both structure and sequence space. For redesigning protein-protein interactions, if a crystal structure of the protein complex being redesigned is available, it is often assumed that the backbone of the redesigned complex will not be perturbed by sequence. Such a fixed-backbone design approach therefore considers only structural degrees of freedoms for the side chains. The repacking of side-chain conformations can be further simplified by sampling from a pre-defined rotamer library[83]. Energy minimization can be used to relieve serious steric clashes that stem from artifacts in discretizing the side-chain conformations.

Although success has been reported for many design applications using a fixed-backbone approach, it is clear that even a small number of mutations can sometimes introduce significant variation in backbone geometry[84,85]. Designing on a fixed backbone therefore risks the elimination of viable sequences that could be otherwise accommodated if backbone flexibility were treated. This is especially relevant for designing specificity, as interaction modeled with an off-target on a fixed backbone might not represent its lowest energy conformation. Different approaches have been proposed to introduce backbone flexibility on a local or global scale[86,87,88,89,90,91,92]. For example, Fu et al. demonstrated that designing on an ensemble of helices generated from normal mode analysis for binding the protein Bcl-xL produced binders with more diversified sequences[89]. Smith et al. found that the incorporation of "backrub sampling", a sampling method inspired by examining small, local structural variations within the

PDB, significantly improved the performance in predicting binding profiles for different PDZ domains[91].

Both deterministic algorithms (e.g. dead end elimination (DEE), A* and integer linear programming) or stochastic ones (e.g. Monte Carlo (MC) and genetic algorithms) can be used for optimization in the structure/sequence space[93,94,95,96,97,98]. Deterministic algorithms are powerful but can present problems when the scoring function contains a non-pairwise decomposable term such as continuum electrostatics, or when backbone flexibility is treated explicitly. Different efforts have been presented to partially overcome such difficulties[99,100,101]. Although stochastic algorithms might not always be able to converge on the optimal solution, they can be more robust in accommodating different formulations/objectives and might be the only viable option when the search space is too large for deterministic methods[98]. Different heuristics have been presented to manage the search problem. For example, search using a fast but less accurate pairwise decomposable scoring function can be performed first to narrow down the sequence space, and a more sophisticated scoring function can then be used for evaluation[102]. Recently, Grigoryan et al. proposed a novel framework, CLASSY, in which a technique called cluster expansion is first used to approximate a structured-based scoring function by a linear sequence-based scoring function[103]; the optimization algorithm integer linear programming (ILP) can then be run for optimization in sequence space only[80]. In addition to dramatically reducing the time spent on evaluating a sequence during the optimization, the ILP formulation allows the incorporation of multiple linear constraints, making it ideal for exploring different trade-offs in a multi-specificity design problem.

Application to designing protein-protein interaction specificity

Successful examples of the computational protein design of interaction specificity have been reported for a number of different proteins. Many of these consisted of redesigning sequences for two interaction partners to create obligate heterodimers or orthogonal protein interfaces[81,102,104,105,106]. Among one of the first examples in explicitly designing for specificity, Havranek et al. designed homo and hetero-specific dimeric coiled coils with novel specificity determinants not present in native coiled-coil sequences[107]. Bolon et al. redesigned the SspB homodimer into an obligate heterodimer, and demonstrated experimentally the importance of explicit negative design in this example[81]. Green et al.[102], Kortemme et al.[104] and Sammond et al.[105] also explored the redesign of native protein interfaces to create designed interfaces that are orthogonal. Potapov et al. presented an interesting approach for such interface redesign by considering a protein interface to be made up of different modules (sets of interconnected residues) independent from one another. A module at the interface between TEM1 β-lactamase and its inhibitor protein BLIP was replaced with another module from an unrelated protein interface. The resulting interface was shown to be orthogonal to the original one and still retained high affinity[106].

Other studies concentrated on redesigning proteins to selectively bind the desired target over a number of undesired off-targets[33,80,108,109]. Yosef et al. reported the redesign of calmodulin to specifically target one peptide sequence over another. Interestingly, only positive design for binding the target was considered, yet the design was verified experimentally to be ~300 fold specific[109]. Continuing the previous discussion on objectives for protein design, it was suggested that explicit negative design might not be necessary in this case when the target and the off-target are significantly different from one another[17]. In a landmark example of explicitly considering

negative design against off-targets for inhibitor design, Grigoryan et al. designed specific inhibitor peptide against all 20 human bZIP families, and subsequent experimental testing verified that some of the designed peptides indeed showed the desired global specificity[80].

Challenges for computational protein design

Significant challenges remain for computational protein design of interaction specificity. These challenges are present in all aspects of protein design. One fundamental limitation is that predicting protein-protein interaction specificity reliably is still a difficult task. Possible sources of deficiencies in physics based models have been suggested in different studies, in addition to the ones described before regarding hydrophobic sidechain packing preferences and hydrogen bonding (see the section on scoring functions). Favorable electrostatics at the protein interface might not be properly balanced with the energy cost of interfacial charge burial[110]. Insufficient structural sampling could produce artificial steric clashes or fail to identify the optimal conformation[90, 111]. These issues can be rather subtle and difficult to improve upon. Various modeling suites that adjust relative weights of different physics based scoring terms or use different approaches for structural sampling have been developed[92,112,113,114], partly guided and tested by available mutational free energy change data. Predictions made from these models correlated with the mutational data to a certain extent, but the agreements were not impressive and their performances on protein-protein interaction specificity could be further compromised. As described before, statistical potentials can possibly address some of the deficiencies in physics based models. However, they present their own deficiencies and have not been demonstrated to be able to make reliable on a global scale either.

Of course the imperfectness of scoring functions should not prevent attempts at computational protein design. In design, researchers enjoy the advantage of testing only sequences predicted to

be optimal, allowing a greater tolerance of prediction error. Researchers can also focus on testing designs generated with strategies that they have higher confidence in. For example, Lippow et al. successfully improved the affinities for different antibodies by mainly optimizing energy contributions from electrostatics[116]. Sammond et al. also presented a series of filters based on general knowledge of protein-protein interactions that the predicted affinity-enhancing mutations need to pass before being tested[117]. However, one could imagine that as the design problem becomes more and more challenging, e.g. a specificity design problem involving multiple off-targets, demands on the accuracy of the scoring function will increase as well, possibly offsetting other advantages offered by computational protein design.

One potential strategy to address deficiencies in physics or statistical based structural modeling is to supplement them with models derived from other sources such as experimental data[118,119]. This was illustrated in the study by Grigoryan et al. on comparing performances of different models for predicting interaction specificity among ~50 human bZIP coiled coils[78]. Two models, one trained by support vector machine (SVM) on independent sequence and experimental data for coiled-coil interactions[120], and another parameterized directly using experimentally measured coupling energies for bZIP coiled coils[121] showed significantly better performances than pure structural models on predicting specificity. One caveat for experimentally derived models is that information not present in the experimental dataset from which the models are derived cannot be learned. Hybrid models that combine physics based structural models with the experimentally derived ones have been developed to address this. These hybrid models were in fact used in the study of Grigoryan et al., as described above, to design globally specific inhibitor peptide against 20 human bZIP family proteins[80]. The accuracy offered by the experimentally derived models is likely crucial for the success of the designs. It is

tempting to generalize such approaches to other interaction specificity design applications as well. However, this approach requires the presence of a large amount of experimental data relevant to the protein being studied, which is often not available for many proteins of interest. Information encoded within the evolutionary history can also be utilized to provide insight on protein-protein interaction specificity[122,123], but such methods often required ample sequencing information and at least some knowledge of interaction patterns across different species for the proteins of interest.

Experimental library screening

Like computational protein design, experimental library screening/selection is motivated by the desire to search among a large number of sequences for ones with the desired properties[124]. Below I first briefly review several key experimental aspects of library screening/selection, including different techniques for generating sequence diversity and different screening/selection platforms. I then focus on the emerging trend of how library screening can be combined with computational protein design to facilitate the discovery of desired sequences.

Generating sequence diversity

The first task in performing a library screen is to generate an ensemble of sequences experimentally. This is typically performed at the DNA level, with diversity translated to protein sequences at a later stage. Genes with randomization at selected positions can be made by PCR-based assembly procedures using partially randomized oligoneucleotides containing degenerate codons[125]. It is possible to encode randomness at a position by using a mixture of multiple non-randomized oligonucletides[126] or by utilizing trinucleotide synthesis[127] as well. However, issues such as how distant the randomized positions are from each other in sequence, how many

randomized positions are placed closely together, and how much diversity needs to be introduced at these positions affect whether the PCR assembly procedure can be performed cost-effectively. Alternatively, diversity can be generated by recombining fragments from different native or synthetic genes, mimicking the process of homologous recombination[128]. It should be noted that common methods for introducing diversity often place constraints on the types of sequences that can be generated. For example, randomization at selected positions dictates that the sequences will be combinatorial with respect to diversity at these positions. Recombination among gene fragments, on the other hand, will result in combinatorial sequences with respect to the fragments. Instead of introducing randomization at defined positions/fragments, mutations can be introduced randomly across the whole protein (or a selected sub-region) using error-prone PCR or other methods[129,130].

Note that different randomization strategies can be combined together. One good example is the process leading to diversification of human antibodies[131]. The variable region of each class of antibody chain is assembled from different types of gene segments (the V, D, and J segment) in a site-directed recombination event known as V-D-J joining. The presence of different variants for each type of gene segment leads to a combinatorial diversity estimated to be bigger than 10^5. Mechanisms such as somatic hypermutation and gene conversion are employed to further increase diversity and generate antibodies of sufficient affinities for their targets (affinity maturation). This process of generating a preliminary pool of sequence diversity from which selected sequences are further optimized can be mimicked *in vitro* as well. For example, randomization can first be introduced in a guided manner by recombining different native or synthetic gene fragments (analogous to V-D-J joining) or by mutating selected positions in a

combinatorial manner. Promising sequences can be identified and techniques like error-prone PCR (analogous to somatic hypermutation) can be performed to further optimize their properties.

Screening/selection platform

Next an appropriate screening/selection platform needs to be chosen. This depends on a number of different factors, including the size of the DNA library and the desired activity to be screened or selected. Different molecular display technologies[132] such as phage display[133], bacterial display[134], yeast display[135], mRNA display[136] and ribosome display[137] have been widely used to screen for desired protein-protein interactions, although other platforms such as yeast two-hybrid have been considered as well. Cell free display methods (mRNA, ribosome) and phage display allow the handling of much larger library sizes ($> 10^{14}$) than those affordable by cell-based display methods (10^{9}-10^{10} for bacterial display and 10^{7}-10^{9} for yeast display). However, for bacterial and yeast display, florescence activated cell sorting (FACS) can be used to sort cells displaying the desired sequences in solution. This bypasses the need to first immobilize and then elute the desired clones from a surface, which is often required in cell-free display technologies. This is advantageous when selecting for protein-protein interaction specificity, as conditions for competition or negative selection can be more easily tuned by simply varying the concentration of target and off-targets.

Combining computational protein design and experimental library screening

Compared to computational protein design, experiments are directly used to evaluate a sequence in library screening, largely removing the worry that the designed sequences might not behave as predicted. However, relative computational protein design, library screening can only be performed for a much more limited number of sequences. This raises concerns as to whether the combinatorial sequence space can be adequately sampled, because mutations picked

randomly are rarely beneficial for the desired trait. Following the thinking behind computational protein design, it is therefore tempting to combine the advantages of these two methods. Instead of computationally designing a few sequences, one can computationally design a library. Sequences in the library will not be chosen randomly, but are instead selected on the basis of computational structural modeling. Although the computational models might not be perfect, as described previously, they could nevertheless help bias the experimental search to a more productive sequence space.

The idea of combining computational protein design and experimental library screening has been explored by several different groups[138-149]. Computationally designing a library presents distinct challenges from designing a selected number of sequences. One major difference is that practical aspects of the chosen experimental strategies need to be given due consideration during the computational design phase. As described before, for most experimental library construction protocols, the diversity of library sequences will be combinatorial with respect to positions or gene fragments. The screening platform also places a limit on the number of sequences that can be tested. Another important difference is that the library design objective is no longer obvious; one must decide whether the designed library should cover a particular set of sequences, whether the predicted behaviors for all library sequences should be evaluated, or if diversity among library sequences is most important. There may be multiple objectives that one aims to consider/optimize, and different trade-offs could exist among these. For example, the desire to create a library at lower cost could mean more restrictions on the type of sequences that could be included. Aiming to sample a larger sequence space through the library could imply a less than ideal coverage for the designed sequences overall.

Below I review the different approaches used to computationally design a library and their applications for different design problems. Although few studies focused on the question of designing protein-protein interaction specificity, the general concepts should have broad relevance.

Designing a library with selected positions randomized

Several approaches have been suggested for designing a protein library with selected positions randomized. In the first approach, a library "score" is first defined, and the library is designed to optimize this score. Treynor et al. defined the score to be the arithmetic average of the energies calculated by structural models of all sequences in the library[140]. Optimization of this library energy is analogous to optimizing the energy of a single sequence, with the search being performed in the space of degenerate codons (could be viewed as sets of amino acids) instead of amino acids. Libraries of green fluorescent protein (GFP) variants were designed accordingly to enrich sequences compatible with the structural fold of GFP. It was observed that these libraries contained a greater fraction of proteins that fluoresced, as well as a greater diversity of colors, compared to an error-prone PCR library. In a separate computational study, Parker et al. also defined library quality to be the averaged energies of all library sequences, but proposed to considered as well another objective that represented the novelty of the library sequences (e.g. when compared to a multiple sequence alignment of homologous native proteins)[148]. Optimization was performed using integer programming in the space of degenerate codons that included consideration of library size. Parker et al. further demonstrated for a few protein systems the trade-off between quality and novelty as defined above computationally. As described before, the concept of trade-offs among different library design considerations is an

important one and worth contemplating when weighing the relative importance of different aspects for library design.

In the second approach, computational protein design is first performed to obtain an ensemble of sequences. An amino-acid profile (i.e. the frequency of different amino acids at each designed position) is derived from these sequences, and the library is designed with the aim to match the diversity observed in the profile. One caveat is that the library obtained accordingly may not reflect the original ensemble of designed sequences. Hayes et al. used this approach to design a library of TEM β-lactamase variants to screen for clones with improved resistance toward the antibiotic cefotaxime[126]. Randomization was introduced to the active site, and compatibility with the protein fold (i.e. the crystal structure of TEM β-lactamase) was assessed in designing the sequence ensemble. Variants with a 1,280 fold increase in resistance were identified out of a ~200,000 member library. Guntas et al. also used this approach to design a library of variants of the ubiquitin ligase E6AP that binds to the NEDD8-conjugating enzyme Ubc12, a nonnatural partner, and obtained multiple tight binders (K_d < 100 nM) from the screen[144]. Degenerate codons were selected at each position by considering their efficiencies in representing the amino-acid diversity profile and the library size. One interesting observation from this work is that equally good performances were obtained for the designed library enriched in predicted binders and that enriched simply in well-folded sequences. Both libraries performed better than a random library.

In the third approach, an amino acid diversity profile is derived from a probabilistic framework rather than from an ensemble of designed sequences. Voigt et al. applied mean-field theory to capture the structural tolerance of each designed position[138]. The study aimed to use the metric as a guide in selecting positions for randomization, and good agreement was observed

with prior experimental directed evolution studies of subtilisin E and T4 lysozyme. Saven and coworkers also proposed using a statistical theory for the design of combinatorial libraries[149,150].

Designing a library made by combining different gene fragments

One challenge in making a library generated by *in vitro* recombination among homologous native or synthetic genes is how to select the cross-over points. Voigt et al. presented SCHEMA, an algorithm to help address this issue[139]. Points for cross-over were chosen so as to minimize disruption of important residue-reside interactions as observed in the crystal structure. The argument is that hybrid proteins generated this way are more likely to be folded and functional. Correlation between cross-over points predicted from SCHEMA and prior *in vitro* recombination experiments was observed, and SCHEMA was subsequently applied to a series of different design problems[151,152].

One advantage for making a library by combining gene fragments is that correlations among residues within the same fragment can be preserved. This is in contrast to designed libraries combinatorial in positions. Although correlation between different designed positions could still play a role in this type of library by deciding what pair of degenerate codons (or sets of amino acids) to choose at two separate positions, no coupling between positions is enforced for the library sequences. As an example of recombining synthetic gene fragments obtained from computational protein design (as opposed to fragments derived from native genes as performed by Voigt et al.), Lippow et al. redesigned a galactose oxidase enzyme to process glucose instead[145]. An ensemble of > 2,000 sequences was first designed computationally. The 12 designed positions were then grouped into 4 assembly regions guided by proximity in sequence. Each region was then encoded by a mixture of synthetic oligonucleotides such that correlations between different positions in each region were preserved. The library was assembled from

these fragments. Using this approach, Lippow et al. successfully identified a variant with 400-fold improvement in activity toward glucose from a 10,000 member library.

Improving computational designs by library screening

One other approach for combining computational protein design and experimental library screening is to use library screening to further improve existing designs. Although computational prediction can be used to guide library design, any prediction model not entirely accurate will generate some bias, meaning that some of the important sequence space will not be sampled, even by a large library. This is especially relevant for difficult prediction/design problems. In this case, a more random/less guided strategy could prove beneficial in identifying important sequence features that would be missed by the model. The approach is demonstrated by Khersonsky et al. for optimizing the catalytic activity of the *in silico* designed Kemp eliminase. Error-prone PCR, gene shuffling and site directed randomization were all employed to generate diverse library sequences derived from the initial computationally designed sequence. Mutants with improvement of > 400 fold in catalytic activity were identified from the screen/selection.

One observation from the studies described above is that the metric used for selecting sequences to be included into the library could be different from the real objective. As described before, Hayes et al.[126], Treynor et al.[140] and Guntas et al.[144] all designed library sequences to be compatible with a structural fold but screened the libraries for function (improved enzymatic activity toward an antibiotic, different photophysical properties, and protein-protein interaction, respectively). Although a well folded sequence is likely a necessary but not sufficient criterion for the presence of different functions, its prediction could be much easier and may in fact represent a more efficient use of the computational prediction model. This could have general implications for difficult design goals such as protein-protein interaction specificity.

Contributions of this thesis in designing protein interaction specificity

One of the major challenges for computational protein design is that current structural models that describe the relationship between sequence and structure (or function) are not always reliable. Above, I summarized and discussed two approaches to address this based on prior studies. The first approach aims to supplement structural models with more restricted but possibly more accurate models derived from experimental data. The second approach utilizes the ability of experimental library screening to survey a more diverse collection of designed sequences, to compensate for the deficiencies of the prediction models.

In this thesis I first describe the design of an inhibitor peptide that binds to the viral bZIP protein BZLF1 (Chapter 2). Specificity against design self-interaction was important for this study. The protein family being studied represents a case for which semi-accurate scoring functions derived from the first approach were available, and successful inhibitor peptides were designed accordingly. In Chapter 3, following the rationale behind the second approach, I present a novel framework that can be used to design libraries to be screened for protein-protein interaction specificity. I applied the framework to the identification of anti-apoptotic Bcl-2 protein variants with novel interaction specificities toward different BH3 peptides. Designed libraries were subjected to screening, and clones with the desired binding specificity were obtained. In Chapter 4, I focus on the question of BH3 sequence determinants of specificity against different anti-apoptotic Bcl-2 proteins. The study focuses on the application of a purely experimental scoring model derived from SPOT arrays to the prediction and design problem. I conclude with a discussion of how the two approaches can be combined together to advance our understanding and our ability to design novel protein-protein interaction specificity.

References

1. Pawson, T. & Nash, P. (2000). Protein-protein interactions define specificity in signal transduction. *Genes Dev* **14**, 1027-47.
2. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74.
3. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-6.
4. Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P. & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8.
5. Newman, J. R. & Keating, A. E. (2003). Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**, 2097-101.
6. Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A. & MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **317**, 364-9.
7. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-30.
8. Lee, H. J. & Zheng, J. J. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun Signal* **8**, 8.
9. Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D. & Sidhu, S. S. (2008). A specificity map for the PDZ domain family. *PLoS Biol* **6**, e239.
10. Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R. & Oschkinat, H. (2004). Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol* **343**, 703-18.
11. Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J. & et al. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell* **72**, 767-78.
12. Liu, B. A., Jablonowski, K., Shah, E. E., Engelmann, B. W., Jones, R. B. & Nash, P. D. SH2 domains recognize contextual peptide sequence information to determine selectivity. *Mol Cell Proteomics* **9**, 2391-404.
13. Zarrinpar, A., Park, S. H. & Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676-80.

14. Tonikian, R., Xin, X., Toret, C. P., Gfeller, D., Landgraf, C., Panni, S., Paoluzi, S., Castagnoli, L., Currell, B., Seshagiri, S., Yu, H., Winsor, B., Vidal, M., Gerstein, M. B., Bader, G. D., Volkmer, R., Cesareni, G., Drubin, D. G., Kim, P. M., Sidhu, S. S. & Boone, C. (2009). Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* **7**, e1000218.
15. Chen, L., Willis, S. N., Wei, A., Smith, B. J., Fletcher, J. I., Hinds, M. G., Colman, P. M., Day, C. L., Adams, J. M. & Huang, D. C. (2005). Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol Cell* **17**, 393-403.
16. Wojtowicz, W. M., Wu, W., Andre, I., Qian, B., Baker, D. & Zipursky, S. L. (2007). A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. *Cell* **130**, 1134-45.
17. Graveley, B. R., Kaur, A., Gunning, D., Zipursky, S. L., Rowen, L. & Clemens, J. C. (2004). The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA* **10**, 1499-506.
18. Schreiber, G. & Keating, A. E. (2011). Protein binding specificity versus promiscuity. *Curr Opin Struct Biol* **21**, 50-61.
19. Kaneko, T., Sidhu, S. S. & Li, S. S. (2011). Evolving specificity from variability for protein interaction domains. *Trends Biochem Sci* **36**, 183-90.
20. Kaneko, T., Huang, H., Zhao, B., Li, L., Liu, H., Voss, C. K., Wu, C., Schiller, M. R. & Li, S. S. Loops govern SH2 domain specificity by controlling access to binding pockets. *Sci Signal* **3**, ra34.
21. Appleton, B. A., Zhang, Y., Wu, P., Yin, J. P., Hunziker, W., Skelton, N. J., Sidhu, S. S. & Wiesmann, C. (2006). Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity. *J Biol Chem* **281**, 22312-20.
22. Kimber, M. S., Nachman, J., Cunningham, A. M., Gish, G. D., Pawson, T. & Pai, E. F. (2000). Structural basis for specificity switching of the Src SH2 domain. *Mol Cell* **5**, 1043-9.
23. Al-Lazikani, B., Lesk, A. M. & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* **273**, 927-48.
24. Meenan, N. A., Sharma, A., Fleishman, S. J., Macdonald, C. J., Morel, B., Boetzel, R., Moore, G. R., Baker, D. & Kleanthous, C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc Natl Acad Sci U S A* **107**, 10080-5.
25. Vinson, C., Acharya, A. & Taparowsky, E. J. (2006). Deciphering B-ZIP transcription factor interactions in vitro and in vivo. *Biochim Biophys Acta* **1759**, 4-12.
26. Gretes, M., Lim, D. C., de Castro, L., Jensen, S. E., Kang, S. G., Lee, K. J. & Strynadka, N. C. (2009). Insights into positive and negative requirements for protein-protein interactions by crystallographic analysis of the beta-lactamase inhibitory proteins BLIP, BLIP-I, and BLP. *J Mol Biol* **389**, 289-305.
27. Skelton, N. J., Koehler, M. F., Zobel, K., Wong, W. L., Yeh, S., Pisabarro, M. T., Yin, J. P., Lasky, L. A. & Sidhu, S. S. (2003). Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. *J Biol Chem* **278**, 7645-54.

28. Songyang, Z., Gish, G., Mbamalu, G., Pawson, T. & Cantley, L. C. (1995). A single point mutation switches the specificity of group III Src homology (SH) 2 domains to that of group I SH2 domains. *J Biol Chem* **270**, 26029-32.
29. Weng, Z., Rickles, R. J., Feng, S., Richard, S., Shaw, A. S., Schreiber, S. L. & Brugge, J. S. (1995). Structure-function analysis of SH3 domains: SH3 binding specificity altered by single amino acid substitutions. *Mol Cell Biol* **15**, 5627-34.
30. Li, W., Dennis, C. A., Moore, G. R., James, R. & Kleanthous, C. (1997). Protein-protein interaction specificity of Im9 for the endonuclease toxin colicin E9 defined by homologue-scanning mutagenesis. *J Biol Chem* **272**, 22253-8.
31. Van der Sloot, A. M., Kiel, C., Serrano, L. & Stricher, F. (2009). Protein design in biological networks: from manipulating the input to modifying the output. *Protein Eng Des Sel* **22**, 537-42.
32. Olive, M., Williams, S. C., Dezan, C., Johnson, P. F. & Vinson, C. (1996). Design of a C/EBP-specific, dominant-negative bZIP protein with both inhibitory and gain-of-function properties. *J Biol Chem* **271**, 2040-7.
33. van der Sloot, A. M., Tur, V., Szegezdi, E., Mullally, M. M., Cool, R. H., Samali, A., Serrano, L. & Quax, W. J. (2006). Designed tumor necrosis factor-related apoptosis-inducing ligand variants initiating apoptosis exclusively via the DR5 receptor. *Proc Natl Acad Sci U S A* **103**, 8634-9.
34. Lee, E. F., Czabotar, P. E., van Delft, M. F., Michalak, E. M., Boyle, M. J., Willis, S. N., Puthalakath, H., Bouillet, P., Colman, P. M., Huang, D. C. & Fairlie, W. D. (2008). A novel BH3 ligand that selectively targets Mcl-1 reveals that apoptosis can proceed without Mcl-1 degradation. *J Cell Biol* **180**, 341-55.
35. Bolon, D. N., Wah, D. A., Hersch, G. L., Baker, T. A. & Sauer, R. T. (2004). Bivalent tethering of SspB to ClpXP is required for efficient substrate delivery: a protein-design study. *Mol Cell* **13**, 443-9.
36. Szczepek, M., Brondani, V., Buchel, J., Serrano, L., Segal, D. J. & Cathomen, T. (2007). Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nat Biotechnol* **25**, 786-93.
37. Czyzyk, J., Brogdon, J. L., Badou, A., Henegariu, O., Preston Hurlburt, P., Flavell, R. & Bottomly, K. (2003). Activation of CD4 T cells by Raf-independent effectors of Ras. *Proc Natl Acad Sci U S A* **100**, 6003-8.
38. Dreze, M., Charloteaux, B., Milstein, S., Vidalain, P. O., Yildirim, M. A., Zhong, Q., Svrzikapa, N., Romero, V., Laloux, G., Brasseur, R., Vandenhaute, J., Boxem, M., Cusick, M. E., Hill, D. E. & Vidal, M. (2009). 'Edgetic' perturbation of a C. elegans BCL2 ortholog. *Nat Methods* **6**, 843-9.
39. Bashor, C. J., Helman, N. C., Yan, S. & Lim, W. A. (2008). Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. *Science* **319**, 1539-43.
40. Kiel, C., Yus, E. & Serrano, L. (2010). Engineering signal transduction pathways. *Cell* **140**, 33-47.
41. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20.
42. Cunningham, B. C. & Wells, J. A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081-5.

43. Pons, J., Rajpal, A. & Kirsch, J. F. (1999). Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. *Protein Sci* **8**, 958-68.
44. Boersma, M. D., Sadowsky, J. D., Tomita, Y. A. & Gellman, S. H. (2008). Hydrophile scanning as a complement to alanine scanning for exploring and manipulating protein-protein recognition: application to the Bim BH3 domain. *Protein Sci* **17**, 1232-40.
45. Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**, 775-91.
46. Butterfoss, G. L. & Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu Rev Biophys Biomol Struct* **35**, 49-65.
47. Lippow, S. M. & Tidor, B. (2007). Progress in computational protein design. *Curr Opin Biotechnol* **18**, 305-11.
48. Karanicolas, J. & Kuhlman, B. (2009). Computational design of affinity and specificity at protein-protein interfaces. *Curr Opin Struct Biol* **19**, 458-63.
49. Mandell, D. J. & Kortemme, T. (2009). Computer-aided design of functional protein interactions. *Nat Chem Biol* **5**, 797-807.
50. Boas, F. E. & Harbury, P. B. (2007). Potential energy functions for protein design. *Curr Opin Struct Biol* **17**, 199-204.
51. Gogonea, V., Suarez, D., van der Vaart, A. & Merz, K. M., Jr. (2001). New developments in applying quantum mechanics to proteins. *Curr Opin Struct Biol* **11**, 217-23.
52. Senn, H. M. & Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew Chem Int Ed Engl* **48**, 1198-229.
53. Ponder, J. W. & Case, D. A. (2003). Force fields for protein simulations. *Adv Protein Chem* **66**, 27-85.
54. Roux, B. & Simonson, T. (1999). Implicit solvent models. *Biophys Chem* **78**, 1-20.
55. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science* **268**, 1144-9.
56. Bashford, D. & Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem* **51**, 129-52.
57. Green, D. F. & Tidor, B. (2003). Evaluation of electrostatic interactions. *Curr Protoc Bioinformatics* **Chapter 8**, Unit 8 3.
58. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199-203.
59. Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci U S A* **84**, 3086-90.
60. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133-52.
61. Levy, R. M., Zhang, L. Y., Gallicchio, E. & Felts, A. K. (2003). On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. *J Am Chem Soc* **125**, 9523-30.
62. Chen, J. & Brooks, C. L., 3rd. (2008). Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys Chem Chem Phys* **10**, 471-81.

63. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**, 623-44.
64. Zhou, H. & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11**, 2714-26.
65. Shen, M. Y. & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507-24.
66. Carter, C. W., Jr., LeFebvre, B. C., Cammer, S. A., Tropsha, A. & Edgell, M. H. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol* **311**, 625-38.
67. Feng, Y., Kloczkowski, A. & Jernigan, R. L. (2007). Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins* **68**, 57-66.
68. DeWitte, R. S. & Shakhnovich, E. I. (1994). Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Protein Sci* **3**, 1570-81.
69. Lu, M., Dousis, A. D. & Ma, J. (2008). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J Mol Biol* **376**, 288-301
70. Summa, C. M., Levitt, M. & Degrado, W. F. (2005). An atomic environment potential for use in protein structure prediction. *J Mol Biol* **352**, 986-1001.
71. Misura, K. M., Morozov, A. V. & Baker, D. (2004). Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J Mol Biol* **342**, 651-64.
72. Morozov, A. V., Kortemme, T., Tsemekhman, K. & Baker, D. (2004). Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc Natl Acad Sci U S A* **101**, 6946-51.
73. Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol* **257**, 457-69.
74. Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol* **16**, 166-71.
75. Lu, L., Lu, H. & Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350-64.
76. Clark, L. A. & van Vlijmen, H. W. (2008). A knowledge-based forcefield for protein-protein interface design. *Proteins* **70**, 1540-50
77. Das, R. & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu Rev Biochem* **77**, 363-82.
78. Grigoryan, G. & Keating, A. E. (2006). Structure-based prediction of bZIP partnering specificity. *J Mol Biol* **355**, 1125-42.
79. Alvizo, O. & Mayo, S. L. (2008). Evaluating and optimizing computational protein design force fields using fixed composition-based negative design. *Proc Natl Acad Sci U S A* **105**, 12242-7.
80. Grigoryan, G., Reinke, A. W. & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859-64.
81. Bolon, D. N., Grant, R. A., Baker, T. A. & Sauer, R. T. (2005). Specificity versus stability in computational protein design. *Proc Natl Acad Sci U S A* **102**, 12724-9.

82. Fromer, M. & Shifman, J. M. (2009). Tradeoff between stability and multispecificity in the design of promiscuous proteins. *PLoS Comput Biol* **5**, e1000627.
83. Dunbrack, R. L., Jr. (2002). Rotamer libraries in the 21st century. *Curr Opin Struct Biol* **12**, 431-40.
84. Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1993). The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science* **262**, 1715-8.
85. Lim, W. A., Hodel, A., Sauer, R. T. & Richards, F. M. (1994). The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci U S A* **91**, 423-7.
86. Mandell, D. J. & Kortemme, T. (2009). Backbone flexibility in computational protein design. *Curr Opin Biotechnol* **20**, 420-8.
87. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science* **282**, 1462-7.
88. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J Mol Biol* **290**, 305-18.
89. Fu, X., Apgar, J. R. & Keating, A. E. (2007). Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-xL. *J Mol Biol* **371**, 1099-117.
90. Smith, C. A. & Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol* **380**, 742-56.
91. Smith, C. A. & Kortemme, T. (2010). Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J Mol Biol* **402**, 460-74.
92. Benedix, A., Becker, C. M., de Groot, B. L., Caflisch, A. & Bockmann, R. A. (2009). Predicting free energy changes using structural ensembles. *Nat Methods* **6**, 3-4.
93. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-42.
94. Desmet, J., Spriet, J. & Lasters, I. (2002). Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**, 31-43.
95. Leach, A. R. & Lemon, A. P. (1998). Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins* **33**, 227-39.
96. Kingsford, C. L., Chazelle, B. & Singh, M. (2005). Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* **21**, 1028-36.
97. Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J Mol Biol* **236**, 918-39.
98. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **299**, 789-803.
99. Georgiev, I., Keedy, D., Richardson, J. S., Richardson, D. C. & Donald, B. R. (2008). Algorithm for backrub motions in protein design. *Bioinformatics* **24**, i196-204.
100. Georgiev, I., Lilien, R. H. & Donald, B. R. (2008). The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *J Comput Chem* **29**, 1527-42.

101. Barth, P., Alber, T. & Harbury, P. B. (2007). Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc Natl Acad Sci U S A* **104**, 4898-903.
102. Green, D. F., Dennis, A. T., Fam, P. S., Tidor, B. & Jasanoff, A. (2006). Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry* **45**, 12547-59.
103. Grigoryan, G., Zhou, F., Lustig, S. R., Ceder, G., Morgan, D. & Keating, A. E. (2006). Ultra-fast evaluation of protein energies directly from sequence. *PLoS Comput Biol* **2**, e63.
104. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nat Struct Mol Biol* **11**, 371-9.
105. Sammond, D. W., Eletr, Z. M., Purbeck, C. & Kuhlman, B. Computational design of second-site suppressor mutations at protein-protein interfaces. *Proteins* **78**, 1055-65.
106. Potapov, V., Reichmann, D., Abramovich, R., Filchtinski, D., Zohar, N., Ben Halevy, D., Edelman, M., Sobolev, V. & Schreiber, G. (2008). Computational redesign of a protein-protein interface for high affinity and binding specificity using modular architecture and naturally occurring template fragments. *J Mol Biol* **384**, 109-19.
107. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat Struct Biol* **10**, 45-52.
108. Barth, P., Schoeffler, A. & Alber, T. (2008). Targeting metastable coiled-coil domains by computational design. *J Am Chem Soc* **130**, 12038-44.
109. Yosef, E., Politi, R., Choi, M. H. & Shifman, J. M. (2009). Computational design of calmodulin mutants with up to 900-fold increase in binding specificity. *J Mol Biol* **385**, 1470-80.
110. Sharabi, O., Dekel, A. & Shifman, J. M. Triathlon for energy functions: who is the winner for design of protein-protein interactions? *Proteins* **79**, 1487-98.
111. Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. Automated minimization of steric clashes in protein structures. *Proteins* **79**, 261-70.
112. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**, 369-87.
113. Pokala, N. & Handel, T. M. (2005). Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J Mol Biol* **347**, 203-27.
114. Yin, S., Ding, F. & Dokholyan, N. V. (2007). Modeling backbone flexibility improves protein stability estimation. *Structure* **15**, 1567-76.
115. Potapov, V., Cohen, M. & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel* **22**, 553-60.
116. Lippow, S. M., Wittrup, K. D. & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol* **25**, 1171-6.
117. Sammond, D. W., Eletr, Z. M., Purbeck, C., Kimple, R. J., Siderovski, D. P. & Kuhlman, B. (2007). Structure-based protocol for identifying mutations that enhance protein-protein binding affinities. *J Mol Biol* **371**, 1392-404.

118. Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A. & MacBeath, G. (2008). Predicting PDZ domain-peptide interactions from primary sequences. *Nat Biotechnol* **26**, 1041-5.
119. Gfeller, D., Butty, F., Wierzbicka, M., Verschueren, E., Vanhee, P., Huang, H., Ernst, A., Dar, N., Stagljar, I., Serrano, L., Sidhu, S. S., Bader, G. D. & Kim, P. M. The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol* **7**, 484
120. Fong, J. H., Keating, A. E. & Singh, M. (2004). Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol* **5**, R11.
121. Acharya, A., Rishi, V. & Vinson, C. (2006). Stability of 100 homo and heterotypic coiled-coil a-a' pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry* **45**, 11324-32.
122. Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M. & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043-54.
123. Ashenberg, O., Rozen-Gagnon, K., Laub, M. T. & Keating, A. E. (2011). Determinants of homodimerization specificity in histidine kinases. *J Mol Biol* **413**, 222-35.
124. Jackel, C., Kast, P. & Hilvert, D. (2008). Protein design by directed evolution. *Annu Rev Biophys* **37**, 153-73.
125. Mena, M. A. & Daugherty, P. S. (2005). Automated design of degenerate codon libraries. *Protein Eng Des Sel* **18**, 559-61.
126. Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A. & Dahiyat, B. I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proc Natl Acad Sci U S A* **99**, 15926-31.
127. Kayushin, A. L., Korosteleva, M. D., Miroshnikov, A. I., Kosch, W., Zubov, D. & Piel, N. (1996). A convenient approach to the synthesis of trinucleotide phosphoramidites--synthons for the generation of oligonucleotide/peptide libraries. *Nucleic Acids Res* **24**, 3748-55.
128. Zhao, H. & Arnold, F. H. (1997). Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res* **25**, 1307-8.
129. Cirino, P. C., Mayer, K. M. & Umeno, D. (2003). Generating mutant libraries using error-prone PCR. *Methods Mol Biol* **231**, 3-9.
130. Shivange, A. V., Marienhagen, J., Mundhada, H., Schenk, A. & Schwaneberg, U. (2009). Advances in generating functional diversity for directed protein evolution. *Curr Opin Chem Biol* **13**, 19-25.
131. Frieder, D., Larijani, M., Tang, E., Parsa, J. Y., Basit, W. & Martin, A. (2006). Antibody diversification: mutational mechanisms and oncogenesis. *Immunol Res* **35**, 75-88.
132. Levin, A. M. & Weiss, G. A. (2006). Optimizing the affinity and specificity of proteins with molecular display. *Mol Biosyst* **2**, 49-57.
133. Sidhu, S. S. & Koide, S. (2007). Phage display for engineering and analyzing protein interaction interfaces. *Curr Opin Struct Biol* **17**, 481-7.
134. Georgiou, G., Poetschke, H. L., Stathopoulos, C. & Francisco, J. A. (1993). Practical applications of engineering gram-negative bacterial cell surfaces. *Trends Biotechnol* **11**, 6-10.

135. Feldhaus, M. J., Siegel, R. W., Opresko, L. K., Coleman, J. R., Feldhaus, J. M., Yeung, Y. A., Cochran, J. R., Heinzelman, P., Colby, D., Swers, J., Graff, C., Wiley, H. S. & Wittrup, K. D. (2003). Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat Biotechnol* **21**, 163-70.
136. Roberts, R. W. & Szostak, J. W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc Natl Acad Sci U S A* **94**, 12297-302.
137. Hanes, J. & Pluckthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* **94**, 4937-42.
138. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. G. (2001). Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci U S A* **98**, 3778-83.
139. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat Struct Biol* **9**, 553-8.
140. Treynor, T. P., Vizcarra, C. L., Nedelcu, D. & Mayo, S. L. (2007). Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc Natl Acad Sci U S A* **104**, 48-53.
141. Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci U S A* **107**, 19838-43.
142. Chica, R. A., Moore, M. M., Allen, B. D. & Mayo, S. L. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc Natl Acad Sci U S A* **107**, 20257-62.
143. Barderas, R., Desmet, J., Timmerman, P., Meloen, R. & Casal, J. I. (2008). Affinity maturation of antibodies assisted by in silico modeling. *Proc Natl Acad Sci U S A* **105**, 9029-34.
144. Guntas, G., Purbeck, C. & Kuhlman, B. Engineering a protein-protein interface using a computationally designed library. *Proc Natl Acad Sci U S A* **107**, 19296-301.
145. Lippow, S. M., Moon, T. S., Basu, S., Yoon, S. H., Li, X., Chapman, B. A., Robison, K., Lipovsek, D. & Prather, K. L. Engineering enzyme specificity using computational design of a defined-sequence library. *Chem Biol* **17**, 1306-15.
146. Saraf, M. C., Moore, G. L., Goodey, N. M., Cao, V. Y., Benkovic, S. J. & Maranas, C. D. (2006). IPRO: an iterative computational protein library redesign and optimization procedure. *Biophys J* **90**, 4167-80.
147. Pantazes, R. J., Saraf, M. C. & Maranas, C. D. (2007). Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. *Protein Eng Des Sel* **20**, 361-73.
148. Parker, A. S., Griswold, K. E. & Bailey-Kellogg, C. Optimization of combinatorial mutagenesis. *J Comput Biol* **18**, 1743-56.
149. Wang, W. & Saven, J. G. (2002). Designing gene libraries from protein profiles for combinatorial protein experiments. *Nucleic Acids Res* **30**, e120.
150. Kono, H. & Saven, J. G. (2001). Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* **306**, 607-28.
151. Otey, C. R., Landwehr, M., Endelman, J. B., Hiraga, K., Bloom, J. D. & Arnold, F. H. (2006). Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol* **4**, e112.

152. Heinzelman, P., Snow, C. D., Smith, M. A., Yu, X., Kannan, A., Boulware, K., Villalobos, A., Govindarajan, S., Minshull, J. & Arnold, F. H. (2009). SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *J Biol Chem* **284**, 26229-33.
153. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-5.
154. Khersonsky, O., Rothlisberger, D., Dym, O., Albeck, S., Jackson, C. J., Baker, D. & Tawfik, D. S. Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. *J Mol Biol* **396**, 1025-42.
155. Khersonsky, O., Rothlisberger, D., Wollacott, A. M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K. N., Baker, D. & Tawfik, D. S. Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *J Mol Biol* **407**, 391-412.
156. Musacchio, A., Saraste, M. & Wilmanns, M. (1994). High-resolution crystal structures of tyrosine kinase SH3 domains complexed with proline-rich peptides. *Nat Struct Biol* **1**, 546-51.
157. Poy, F., Yaffe, M. B., Sayos, J., Saxena, K., Morra, M., Sumegi, J., Cantley, L. C., Terhorst, C. & Eck, M. J. (1999). Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Mol Cell* **4**, 555-61.
158. Birrane, G., Chung, J. & Ladias, J. A. (2003). Novel mode of ligand recognition by the Erbin PDZ domain. *J Biol Chem* **278**, 1399-402.
159. Glover, J. N. & Harrison, S. C. (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**, 257-61.
160. Fire, E., Gulla, S. V., Grant, R. A. & Keating, A. E. Mcl-1-Bim complexes accommodate surprising point mutations via minor structural changes. *Protein Sci* **19**, 507-19.
161. Meijers, R., Puettmann-Holgado, R., Skiniotis, G., Liu, J. H., Walz, T., Wang, J. H. & Schmucker, D. (2007). Structural basis of Dscam isoform specificity. *Nature* **449**, 487-91.

Chapter 2

Design of peptide inhibitors that bind the bZIP domain of Epstein-Barr virus protein BZLF1

Reproduced with permission of Elsevier B.V. from

Chen, T. S., Reinke, A. W. & Keating, A. E. Design of peptide inhibitors that bind the bZIP domain of Epstein-Barr virus protein BZLF1. *J Mol Biol* 408, 304-20

Collaborator notes

Aaron Reinke performed the electrophoretic mobility shift assay

Introduction

The basic-region leucine-zipper (bZIP) transcription factors are a large class of proteins conserved in eukaryotes and several viruses that regulate a wide range of biological processes. The structure of bZIP-DNA complexes is very simple: a helical and positively charged DNA-binding region is contiguous with a coiled coil that mediates protein homo- or hetero-dimerization.[1] The bZIP coiled-coil helices wrap around one another in a parallel orientation with "knobs-into-holes" side-chain packing geometry, and a 7-amino-acid heptad repeat characterizes the structure, in which each residue can be assigned a register position labeled *a* through *g* (Fig. 2.1). High-affinity binding of bZIP transcription factors to DNA requires protein dimerization.

Given the many important biological roles of the bZIPs, molecules that selectively disrupt bZIP-DNA interactions could be valuable reagents and even potential therapeutics. Several strategies have been reported for identifying inhibitors. Small molecules have been discovered via high-throughput screening,[2, 3] and peptides that bind to the coiled-coil regions of the bZIPs and disrupt dimer formation have been selected from targeted combinatorial libraries.[4, 5, 6] A particularly effective strategy for blocking bZIP-DNA interactions was developed by Vinson and co-workers, who created a series of dominant-negative peptide inhibitors by replacing the basic regions of certain bZIP proteins with a sequence enriched in negatively charged residues (the "acidic extension"), giving so-called A-ZIPs.[7, 8, 9, 10] The A-ZIPs bind tightly and selectively to bZIPs and have been used to study the effects of inhibiting dimerization and hence DNA binding in both cell culture and animal models.[11, 12]

Current understanding of bZIP coiled-coil interactions has also enabled the computational design of synthetic peptides to block bZIP dimerization. Significant effort has been dedicated to elucidating sequence determinants governing the interactions of bZIP coiled coils, and to

developing predictive computational models that capture these. Several types of residue-pair interactions that are important for specificity have been characterized in detail over the past 20 years, and models derived from physics-based calculations, machine learning, and experimentally measured coupling energies have been developed to explain and predict bZIP coiled-coil interactions.[4, 13, 14, 15, 16, 17] Using such binding models, Grigoryan et al. recently designed a series of peptides that bind to targets in 19 out of 20 human bZIP families.[18]

An interesting issue in the study of bZIP interactions is specificity. Given the similarities among sequences, and the many bZIPs in most eukaryotes, a large number of homo- and heterodimers can potentially form. Interactions among human bZIPs have been shown to be highly selective when assayed *in vitro*,[19, 20] but it can be difficult to achieve specificity in designed bZIP-like peptides. In particular, peptides engineered to bind to bZIP coiled-coil regions have been shown to self-associate strongly and also interact with undesired partners.[5, 18] In this work we address considerations of both affinity and anti-homodimer specificity in the design of peptide inhibitors for a viral bZIP protein, BZLF1.

BZLF1 (Zta, ZEBRA, EB1) is encoded by the Epstein-Barr virus (EBV) and triggers the virus's latent to lytic switch by functioning as a transcription factor and regulator of DNA replication.[21, 22, 23, 24] Infection by EBV has been linked to several human malignancies such as Hodgkin's disease and Burkitt's lymphoma.[25] The basic region of BZLF1 is highly homologous to that of human bZIPs and is responsible for direct contact with DNA; a coiled-coil region immediately C-terminal to the basic helix mediates dimerization. However, a recent crystal structure and other biochemical studies have revealed several unique features of BZLF1 (Fig. 2.1a).[26, 27] The coiled-coil region at the dimerization interface is only 4 heptads long, whereas the coiled-coil regions of human bZIPs typically contain at least 5 heptads. Furthermore, only one of

the four BZLF1 coiled-coil heptads includes a leucine residue at the *d* position; this residue occurs with much higher frequency in human bZIP sequences (hence the name "leucine zipper"). The stability of the BZLF1 homodimer is significantly enhanced by a unique C-terminal (CT) region that folds back on the coiled coil to form additional contacts;[27] the CT region is only partially observed in the crystal structure. Prior work using peptide arrays showed that BZLF1 constructs corresponding to the coiled coil or the coiled coil plus the CT region homo-associate in preference to binding any of 33 representative human bZIP proteins.[28]

It has been shown that a peptide corresponding to the coiled-coil region of BZLF1, lacking the DNA binding residues, inhibits BZLF1 binding to DNA at high micromolar concentrations.[29] In this work, we sought new peptides that would mimic the coiled-coil interface of the native structure yet provide more potent inhibition of DNA binding. As a design target, BZLF1 is both simpler and more complex than human and viral bZIPs that have been the subjects of previous computational design studies.[18, 28] It is simpler because of its unique structural features, which make coiled-coil inhibitors designed to target it unlikely to interact broadly with other bZIP proteins. However, it is more complex because the CT region and unusually tight helix packing make the interface unlike the dimerization domains of better-understood bZIPs.[26] Here we explore the extent to which previously applied design strategies can be used successfully in the context of BZLF1. Throughout our analyses, we explicitly addressed two design criteria: affinity for BZLF1 and design self-association, which is an undesirable trait for an inhibitor. The best inhibitor incorporated both elements and included modifications of BZLF1 in both the coiled-coil and DNA-binding regions. As assessed using DNA-binding gel-shift assays, this designed peptide was much more potent than one corresponding to the native dimerization domain.

Results

Computational design of a peptide to bind the N-terminal part of the BZLF1 coiled coil

Our goal was to identify variants of the BZLF1 dimerization domain that would function as more effective dominant negative inhibitors of DNA binding. As described in the Introduction, BZLF1 possesses several unique features as a bZIP design target. These include the unconventional, short coiled-coil segment and the CT region. The CT can be divided into the proximal CT (residues 222 - 231) and the less structured distal CT (residues 232 – 246), as shown in Fig. 1b. We began by re-designing the N-terminal two and a half heptads of the BZLF1 coiled coil (residues 191 – 209, Fig 2.2b), because we anticipated that this segment would provide the greatest opportunity to improve affinity and heterodimer specificity over the native sequence. Residues 210 – 221 also form part of the coiled-coil structure, but additionally engage in non-coiled-coil hydrophobic contacts with the proximal CT as observed in the crystal structure (Fig. 2.1a). Thus, in order to maintain this stabilizing interaction, these residues were not changed in the design.

Both the desired design-target heterodimer and the undesired design homodimer were modeled as parallel, blunt ended coiled coils. We used the CLASSY protein-design algorithm to choose residues at 10 sites in the design, optimizing the predicted affinity of the design-target complex.[18] The scoring function used was based on a hybrid model that included both physics-based and experimentally derived terms and is described further in the Methods. The optimal-affinity design, which we call BD_{cc} (BZLF1 design against the coiled-coil region, shown in Fig. 2.2c), was predicted to be hetero-specific. In design energy units the predicted stabilities were as follows: BZLF1 homodimer: -29 kcal/mol, BD_{cc} homodimer: -32 kcal/mol, BZLF1/BD_{cc} heterodimer: -44 kcal/mol. Although the score for the design self-interaction was close to that

for native BZLF1 coiled-coil homodimerization, the score for the design-target interaction was significantly better. Thus, although CLASSY can be used to improve specificity against undesired states as well as affinity for a target,[18] this was predicted not to be necessary in this case.

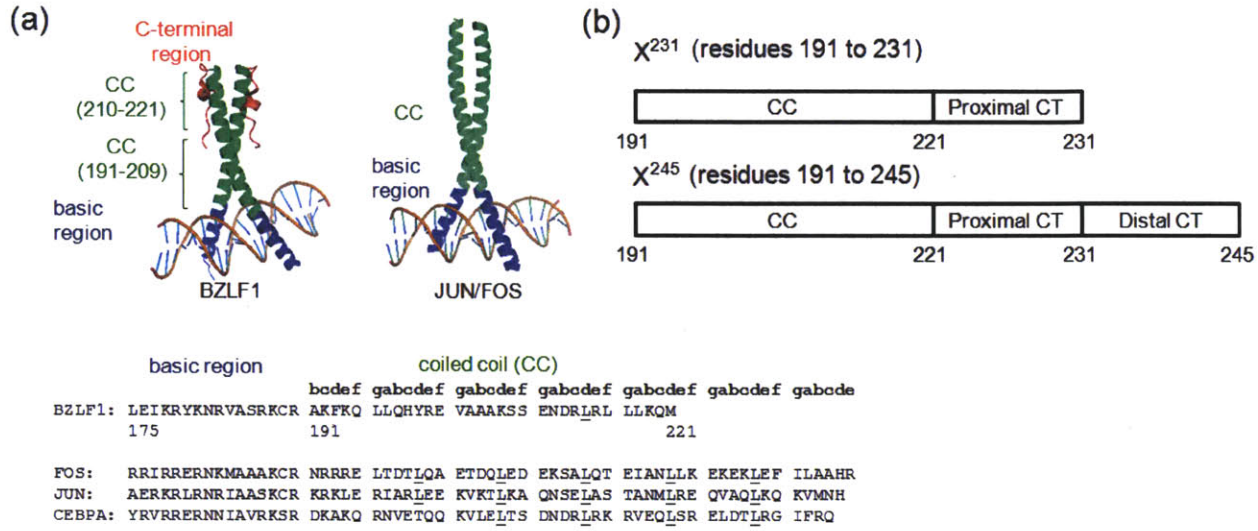


Figure 2.1 Sequence and structure of the BZLF1 bZIP domain.
(a) Crystal structure of BZLF1 bound to DNA26 (PDB ID 2C9L, left) compared to human JUN/FOS bound to DNA54 (PDB ID 1FOS, right). The basic region is blue, the coiled coil is green, and the C-terminal (CT) region is red. At the bottom are sequence alignments for the basic and coiled-coil regions of BZLF1 and representative human bZIPs. Leucines at d positions in the coiled coils are underlined. (b) Scheme of constructs used in this study. The "231" construct includes the coiled coil (CC) and the proximal C-terminal (CT) region; the "245" construct includes the coiled coil (CC) and the full-length C-terminal (CT) region.

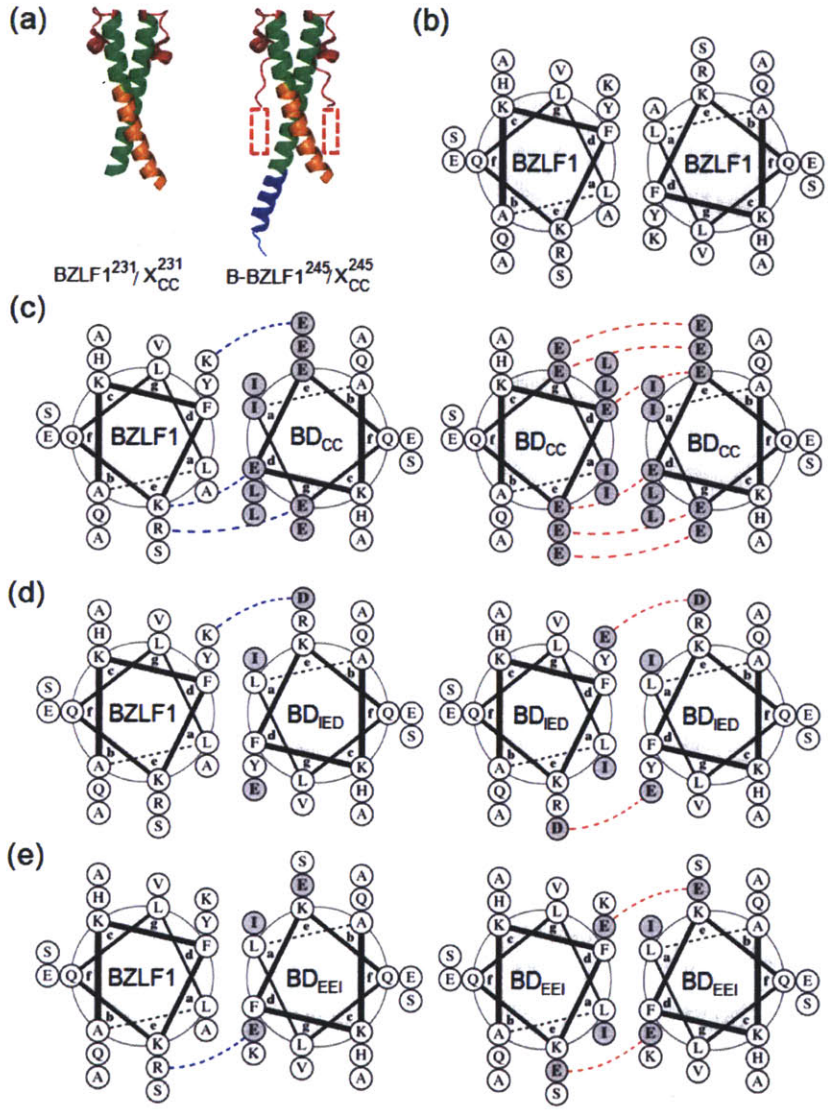


Figure 2.2 Designed inhibitors.

(a) Structural models representing two types of design-BZLF1 complexes tested in this work. At left, the "231" constructs, and at right, the "245" constructs. "X" is a placeholder for the name of a design, e.g. BD_{cc}. Color is as in Fig. 1a except that the designed region is shown in orange. The dashed boxes in the "245" complex indicate that part of the distal CT (237-245) is not resolved in the crystal structure. (b) Helical wheel diagram for the BZLF1 homodimer. (c-e) Helical wheel diagrams for the designs. On the left are design-target heterodimers and on the right are design homodimers. Design residues are highlighted in bold and with a grey background. Potential electrostatic interactions are indicated in blue if attractive and red if repulsive. (c) Design BD_{cc}, (d) Design BD_{IED}, (e) Design BD_{EEI}. In all helical wheel diagrams, only residues from *b* position 191 (Ala) to *f* position 209 (Ser) are shown (this region is orange in Fig. 2.2a), with the helix proceeding from N-to-C terminus into the page. Diagrams generated using DrawCoil 1.0 (http://www.gevorggrigoryan.com/drawcoil/).

The BD_{cc} solution populated most *a* and *d* positions (coiled-coil "core" positions) with Ile and Leu respectively, which are very common in conventional bZIP sequences (Fig. 2.2c). A single *d*-position Glu residue at the extreme N terminus of the coiled coil is uncharacteristic of bZIP sequences, but was predicted to interact favorably with an *e*-position Lys on BZLF1. The five designed *e* and *g* positions (coiled-coil "edge" positions) were all populated with glutamate for improved electrostatic interactions with the target, where three residues in this region are positively charged. Interestingly, predicted charged interactions involved both edge-to-edge (e.g. *g* to *e'*) and core-to-edge (*d* to *e'*) residues in the BZLF1 target, as was previously observed for anti-human bZIP designs.[18] Although core sites occupied by Ile and Leu favor design self-interaction, the charged residues at *e* and *g* are predicted to disfavor it. Charge repulsion is a commonly observed negative design element in many native and model coiled coils.[30, 31, 32, 33]

The anti-BZLF1 peptide was cloned in the context of residues 191- 231 of BZLF1. This construct, BD_{CC}^{231}, includes the entire coiled-coil domain and the proximal CT (Fig. 2.1b, 2.2a, Table 2.1), potentially retaining native interactions observed in the X-ray structure between the C-terminal part of the coiled coil and the CT region. Because the residues optimized in the design calculations are more than 8 Å away from residue I231 in the modeled structure (Fig. 2.2a), the proximal CT excluded from the calculations was not expected to significantly influence the results. Potential interactions between the designed residues and the distal CT, which are not evident in the crystal structure but are suggested by prior studies[27], are addressed in experiments described below.

We used circular dichroism (CD) spectroscopy to study the interaction properties of BD_{CC}^{231}. Thermal denaturation experiments showed that the BD_{CC}^{231} homo-oligomer is destabilized compared to the target homodimer in the same sequence context ($BZLF1^{231}$, residues 191 to 231);

T_m values were 38 °C vs. 43 °C (Fig. 2.3a and Table 2.1). The hetero-complex between BD_{cc}^{231} and $BZLF1^{231}$ (T_m of 53 °C, Table 2.2) was significantly stabilized compared to the $BZLF1^{231}$ homodimer. We conclude that the BD_{cc}^{231} design is very hetero-specific, consistent with expectations based on the design algorithm. The agreement indicates success of the automated CLASSY approach even on a target with a sequence quite different from the human bZIPs.

Designs with weaker self-association

The BD_{cc} design achieved hetero-specificity mostly by improving design-target affinity compared to the native BZLF1 complex. We also sought solutions that achieved hetero-specificity against the same target (the N-terminal part of the BZLF1 coiled coil) by weakening design self-interaction. Toward this end, we tested a negative design strategy that placed charged residues at a core *d* position and the adjacent *e* position such that they would create a local cluster of 4 negative charges in the modeled design coiled-coil homodimer. There are 3 close inter-chain pair contacts in such a cluster (2 *d-e'* interactions and one *d-d'* interaction). We observed variations of this strategy in design solutions obtained using the CLASSY algorithm when optimizing affinity for the target under increasingly stringent constraints limiting the stability of the design homodimer.

We picked two sets of amino-acid changes, (K207E, S208D) and (Y200E, R201E), each corresponding to the (*d, e'*) negative design strategy described above. We also included one stabilizing design element present in the BD_{cc}^{231} solution, A204I (substituting Ile for Ala at an *a* position), to compensate for a potential loss in stability due to the introduction of charge in the core. The resulting two designs were cloned, expressed and purified, again in the context of BZLF1 residues 191 to 231 (204I, 207E, 208D, referred to as BD_{IED}^{231}, and 200E, 201E, 204I,

referred to as BD^{231}_{EEI}, Fig 2.2d-e).

Table 2.1 Sequences[a] and melting temperatures (°C)[b] for BZLF1 and design constructs.

	basic/acid	coiled coil	proximal CT	distal CT	T_m
		191 221 231 245			
		bcdefgabcdefgabcdefgabcdefgabcd			
$BZLF1^{231}$		AKFKQLLQHYREVAAAKSSENDRLRLLLKQM	CPSLDVDSII		43
A-$BZLF1^{231}$	QRAEELARENEELEKEA	EELEQELLKYREVAAAKSSENDRLRLLLKQM	CPSLDVDSII		33
B-$BZLF1^{231}$	LEIKRYKNRVASRKCR	AKFKQLLQHYREVAAAKSSENDRLRLLLKQM	CPSLDVDSII		31
$BZLF1^{245}$		AKFKQLLQHYREVAAAKSSENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	71
A-$BZLF1^{245}$	QRAEELARENEELEKEA	EELEQELLKYREVAAAKSSENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	43
B-$BZLF1^{245}$	LEIKRYKNRVASRKCR	AKFKQLLQHYREVAAAKSSENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	67
BD^{231}_{CC}		AKEEQEIQHLEEEIAALESENDRLRLLLKQM	CPSLDVDSII		38
BD^{245}_{CC}		AKEEQEIQHLEEEIAALESENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	40
A-BD^{245}_{CC}	QRAEELARENEELEKEA	EELEQELLKLEEEIAALESENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	40
BD^{231}_{IED}		AKFKQLLQHYREVIAAEDSENDRLRLLLKQM	CPSLDVDSII		N/A[c]
BD^{245}_{IED}		AKFKQLLQHYREVIAAEDSENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	26
A-BD^{245}_{IED}	QRAEELARENEELEKEA	EELEQELLKYREVIAAEDSENDRLRLLLKQM	CPSLDVDSII	PRTPDVLHEDLLNF	N/A[c]
BD^{231}_{EEI}		AKFKQLLQHEEEVIAAKSSENDRLRLLLKQM	CPSLDVDSII		N/A[c]

[a] The sequences SHHHHHHGESKEYKKGSGS, or GYHHHHHHGSY (the latter for constructs with the acidic extension, A-) should be placed at each N terminus to obtain the full sequences of the recombinant proteins listed in the table. Sites with amino acids different from those of the native sequence (either introduced in the design or as part of the acidic extension) are underlined. Different regions of the sequence (basic region/acidic extension, coiled coil, proximal CT and distal CT) are separated by space. As explained in the text, the acidic extension overlaps the 9 N-terminal residues of the coiled coil. Coiled-coil heptads are indicated using shading.

[b] Total protein concentration was 4 μM.

[c] N/A indicates either lack of cooperative folding or that the observed melting curve indicated the presence of more than one species.

Table 2.2 Melting temperatures (°C) for different BZLF1/design hetero-interactions.

Target	Design	T_m[a]	ΔT_m[b]
$BZLF1^{231}$	BD_{CC}^{231}	53	12 (43/38)
	BD_{IED}^{231}	N/A[c]	N/A[c]
$B\text{-}BZLF1^{245}$	BD_{CC}^{245}	66	12 (67/40)
	BD_{IED}^{245}	N/A[c]	N/A[c]
	$A\text{-}BD_{CC}^{245}$	>80	> 26 (67/40)
	$A\text{-}BZLF1^{245}$	74	19 (67/43)
$B\text{-}BZLF1^{231}$	$A\text{-}BZLF1^{231}$	58	26 (31/33)
JUN	BD_{CC}^{245}	41	10 (23/40)

[a] Total protein concentration was 4 μM.
[b] ΔT_m was obtained by taking the T_m for the hetero-complex and subtracting from it the average of the T_m values for each individual species (listed in parentheses for easy comparison, T_m for the target is shown first, followed by that of the design) when applicable.
[c] N/A indicates either lack of cooperative folding or that the observed melting curve indicated the presence of more than one species.

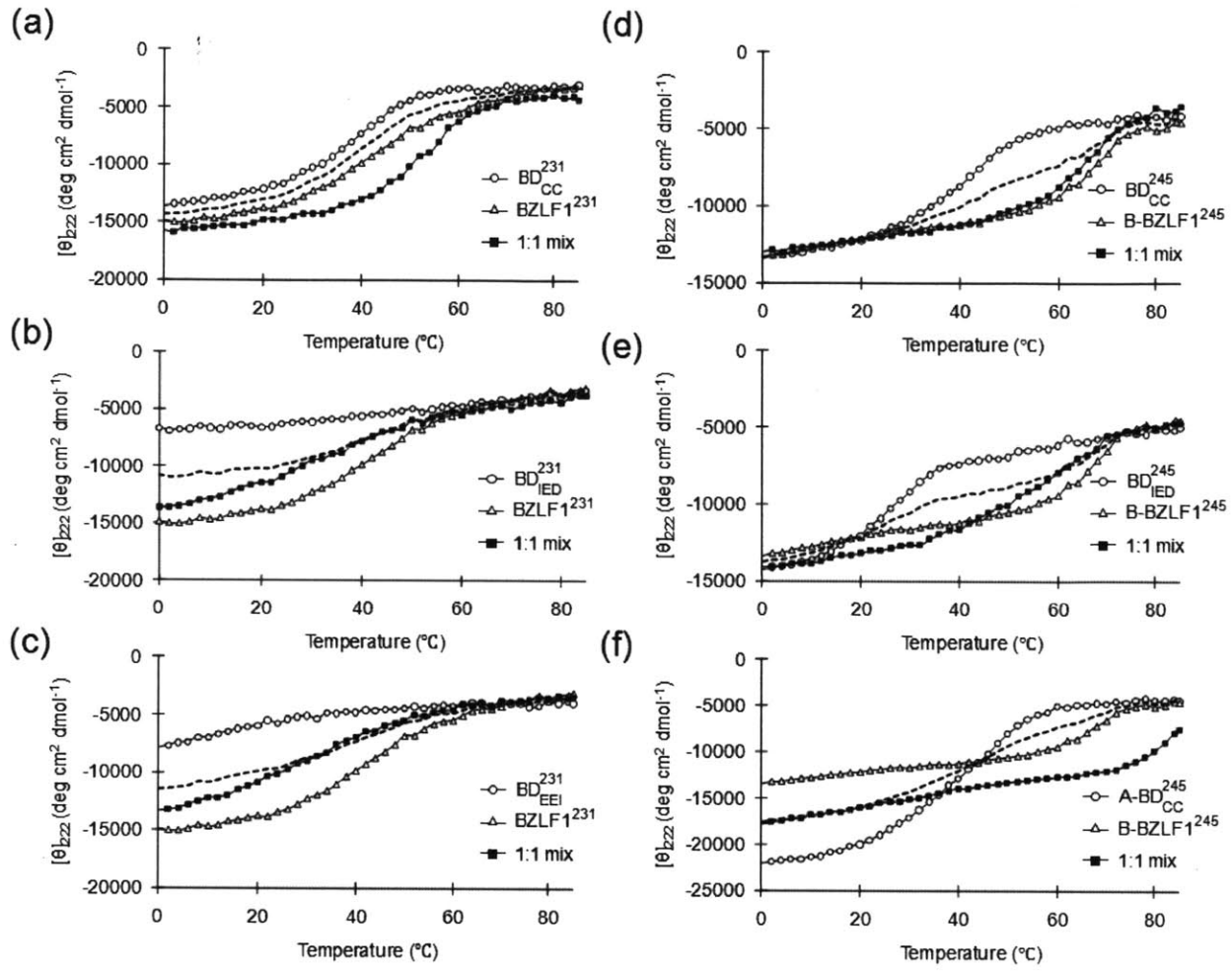


Figure 2.3 Melting curves for targets, designs and complexes monitored by mean residue ellipticity at 222 nm.

Four curves are shown in each panel: the target at 4 μM (open triangles), the design at 4 μM (open circles), a mixture of the target and the design at 2 μM each (closed squares), and the numerical average of the individual melting curves for the target and the design (short dashed lines). The target is $BZLF1^{231}$ for panels (a) - (c) and $B\text{-}BZLF1^{245}$ for panels (d) - (f), as described in text, and the designs are: (a) BD^{231}_{CC}, (b) BD^{231}_{IED}, (c) BD^{231}_{EEI}, (d) BD^{245}_{CC}, (e) BD^{245}_{IED}, and (f) $A\text{-}BD^{245}_{CC}$.

Thermal denaturation experiments monitored by CD showed that both designed peptides, BD^{231}_{IED} and BD^{231}_{EEI}, had relatively weak helical signals even at very low temperatures (Fig. 2.3b, c), illustrating the effectiveness of the negative design strategy. We compared the melting curve for the mixture of each design and $BZLF1^{231}$ with the numerical average of the individual melting

curves for each species (Fig. 2.3b, c). The difference between the two curves below ~22 °C reflects interaction between the designed peptides and $BZLF1^{231}$, and confirms that the designed peptides bind the target more strongly than they interact with themselves. However, an interaction is evident only at low temperatures, indicating that the stability of the design-target complex is lower than the $BZLF1^{231}$ target homodimer. Therefore, these 2 designed peptides represent a specificity profile distinct from that of BD_{CC}^{231}; one that achieves greater destabilization against design self-interaction at the expense of the stability of the design-target interaction.

BD_{CC} and BZLF1 form a heterodimer

We modeled all coiled-coil interactions as parallel, symmetric dimers. Although the oligomerization states of coiled coils can be sensitive to very few amino-acid changes,[34, 35] in BZLF1 the presence of the CT region is expected to strongly favor the parallel dimer geometry observed in the crystal structure for BZLF1. The designed heterodimer also includes an Asn-Asn interaction at *a-a'*, which has been shown to strongly favor dimers, and multiple charged residues at the *e* and *g* positions that are also more prevalent in dimers.[36] Nevertheless, we performed analytical ultracentrifugation (AUC) experiments to study the interaction between BD_{CC}^{231} and $BZLF1^{231}$. Global analysis of sedimentation equilibrium runs performed at multiple concentrations and rotor speeds showed that the best-fit molecular weight for both BD_{CC}^{231} and the 1:1 mixture of BD_{CC}^{231} with $BZLF1^{231}$ corresponded to that expected for a dimer (representative data are shown along with the global fit in Fig. 2.4). For BD_{CC}^{231} with $BZLF1^{231}$, the fitted molecular weight was 104% of that expected for the heterodimer, with a fitted RMS of 0.027 fringes. RMS values obtained by fixing an exact dimer or trimer weight were 0.029 or 0.090 fringes, respectively. For BD_{CC}^{231}, the fitted molecular weight is 102% of that expected for the

homodimer, with a fitted RMS of 0.021 fringes. RMS values obtained by fixing a dimer or a trimer weight were 0.021 or 0.10 fringes, respectively. The AUC data thus confirm the validity of modeling these interactions as dimers.

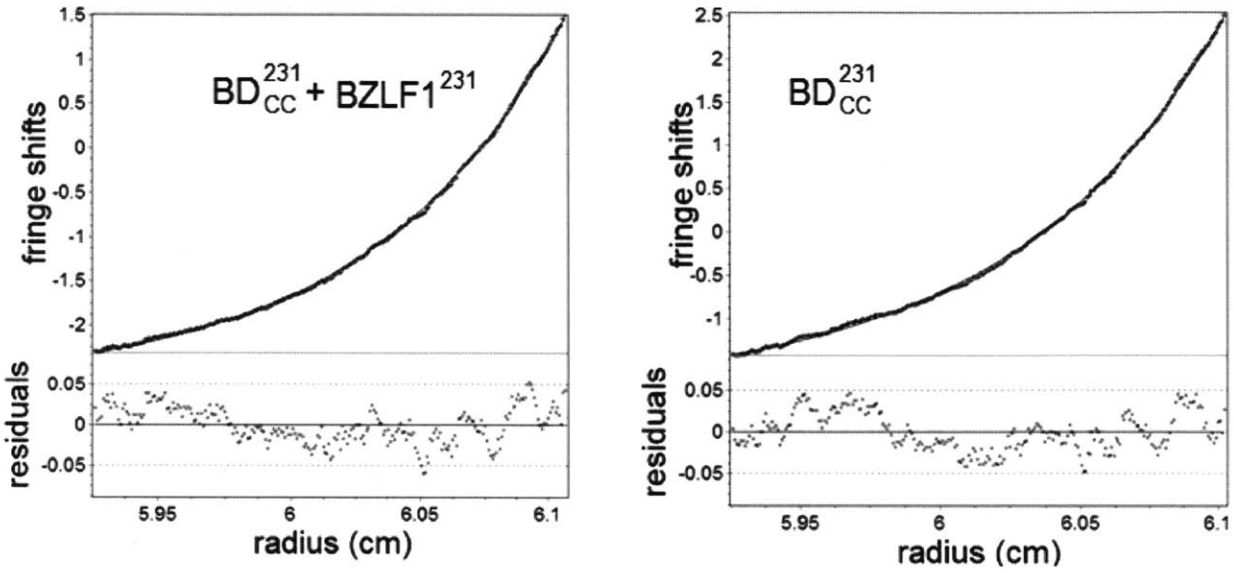


Figure 2.4 Representative analytical ultracentrifugation data for BD_{CC}^{231} + $BZLF1^{231}$ (left) and BD_{CC}^{231} (right).
The fits shown were obtained with data collected at 2 concentrations and 3 different centrifuge speeds. At the bottom are the residuals to the fit.

Testing designs in the full-length BZLF1 dimerization domain

The designs described above targeted the BZLF1 coiled coil and were tested in the context of $BZLF1^{231}$. However, inhibitors of protein function must bind to the full-length protein. One difficulty with designing against the entire BZLF1 dimerization domain (residues 191 - 245) is that the crystal structure shows only the proximal and part of the distal CT region (up to residue 236), with the remaining part of the distal CT region contributing no electron density.[26] Nevertheless, the distal CT region (Fig. 2.1b) has been shown to contribute positively to BZLF1 dimer stability despite possibly being less structured.[27]

We tested whether our design procedures, which considered only the structured coiled coil,

could provide molecules that bind the full-length BZLF1 dimerization domain. For this purpose, a BZLF1 construct that included both the DNA binding basic region and the full-length dimerization domain (termed B-BZLF1^{245}, residues 175-245, Table 2.1) was used instead of BZLF1^{231} as the target. The designed mutations in BD^{231}_{CC} and BD^{231}_{IED} were made in the context of the full-length BZLF1 dimerization domain without the basic region (residues 191-245) to create two new design constructs, BD^{245}_{CC} and BD^{245}_{IED} (Fig. 2.2a, Table 2.1); the distal CT was included in the design constructs to exploit its potentially favorable interaction with the target.

The distal CT dramatically stabilized the BZLF1 homodimer (compare BZLF1^{231} and BZLF1^{245} T_m values of 43 °C and 71 °C, respectively), consistent with prior reports.[27] In contrast, self-association of the BD_{CC} design was not significantly stabilized by the distal CT (Table 2.1). When BD^{245}_{CC} and B-BZLF1^{245} were mixed, there was clear evidence of interaction (Fig. 2.3d, Table 2.2). However, the hetero-interaction between BD^{245}_{CC} and B-BZLF1^{245} did not appear to be stronger than the self-association of the target B-BZLF1^{245} (Table 2.1, 2.2), which contrasts with the behavior of the shorter constructs, BD^{231}_{CC} and BZLF1^{231} (Fig. 2.3a, Table 2.2). Differences in relative stabilities for the shorter and longer constructs suggest that residues in the design do not interact as favorably as the native residues with the distal CT.

In contrast to BD^{245}_{CC}, analysis of BD^{245}_{IED} showed that both the design self-interaction and the design-target interaction were stabilized by the distal CT (compare Fig. 2.3b with Fig. 2.3e). As a result, BD^{245}_{IED} was heterospecific at low temperature. Compared to BD^{245}_{CC}, BD^{245}_{IED} showed weaker self-association but also displayed weaker affinity for B-BZLF1^{245}. Together, the results show that the effect of the distal CT is not negligible and depends on sequence in the coiled-coil region. The impact of the distal CT on the specificity profiles for different designs is considered further in the Discussion.

Specificity of BD_{cc} against human bZIPs

Specificity against human bZIP proteins was not addressed explicitly in our design procedure because we reasoned that the CT region, which is unique to BZLF1, would likely stabilize interaction with BZLF1 but not with human proteins. To assess this, we selected a few human bZIPs and evaluated their interactions with BD_{cc} using CD spectroscopy. To identify those human bZIP proteins most likely to associate with BD_{cc}, we calculated interaction scores with 36 representative human bZIP coiled coils using the scoring function employed in the CLASSY algorithm, which has been shown to be useful for evaluating bZIP coiled-coil associations (Fig. 2.5a).[13,18] Interestingly, BD_{cc} was predicted to interact more favorably with BZLF1 than with any of the human bZIPs, even though the scoring scheme used did not consider interactions involving the CT region. We chose 5 of the top 10 scoring complexes for experimental testing, selecting representative proteins that spanned 5 families and included JUN, the closest predicted competitor. We used constructs for the human proteins that included the basic region and the coiled coil (Fig. 2.5b-f). Analysis of melting curves for each human bZIP and each 1:1 mixture with BD_{cc}^{245} showed that only JUN interacted with BD_{cc}^{245}. The BD_{cc}^{245}/JUN complex, however, was significantly weaker than that between BD_{cc}^{245} and B-$BZLF1^{245}$ (T_m values of 41°C vs. 66°C, Table 2.2). Thus, BD_{cc} is not a promiscuous design and binds preferentially to its target, BZLF1.

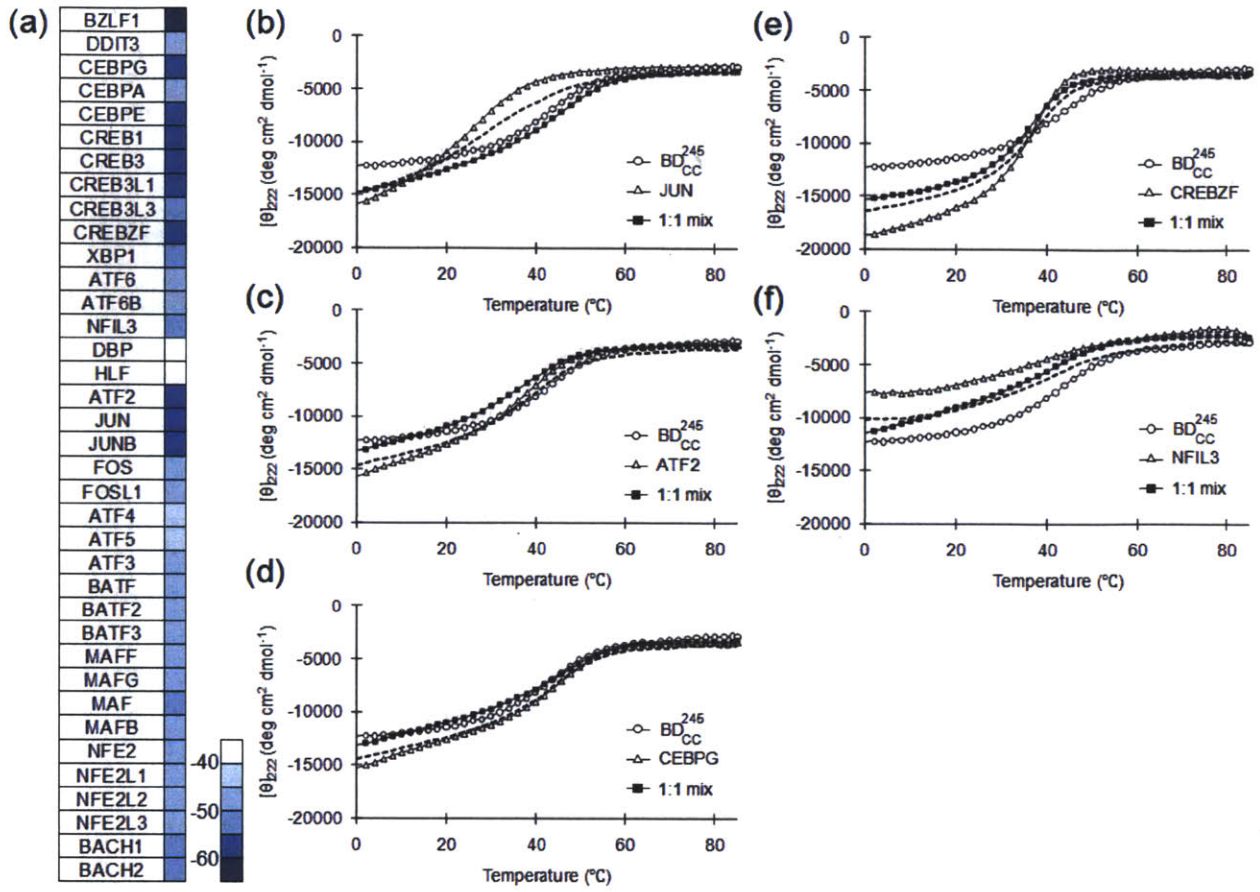


Figure 2.5 Specificity of design against human bZIPs
(a) Predicted scores for BD_{cc} interacting with BZLF1 or human bZIP peptides. (b-f) Melting curves for selected human bZIP peptides, BD^{245}_{CC} or 1:1 mixtures of the two, monitored by mean residue ellipticity at 222 nm. Four curves are shown in each panel: the human bZIP at 4μM (open triangles), BD^{245}_{CC} at 4 μM (open circles), a mixture of the human protein and BD^{245}_{CC} at 2μM each (closed squares), and the numerical average of the individual melting curves for the human bZIP and the design (short dashed lines). The human bZIPs are: (b) JUN, (c) ATF2, (d) CEBPG, (e) CREBZF, and (f) NFIL3.

Enhancing design performance with an N-terminal acidic extension

Vinson and colleagues have shown that replacing the basic region of several native bZIPs with a designed sequence enriched in glutamates can provide potent dominant-negative inhibitors of bZIP dimerization and DNA binding.[7, 9, 10] They also showed that such an acidic extension improved the affinity of a peptide rationally designed to heterodimerize with human

bZIP CEBPA.[8] Because the basic region of BZLF1 is highly similar to that of human bZIPs (Fig. 2.1a), we reasoned that incorporating an acidic extension into the N-terminus of our BD_{CC}^{245} design might enhance its affinity for BZFL1.

Three acidic extension variants developed by Vinson et al. differ in 2 positions that could interact with the BZLF1 basic region, if the interaction occurred with a coiled-coil-like geometry as has been hypothesized for other systems.[7] We chose to use the "A"-extension, which introduced the possibility of an attractive Glu-Arg *g-e'* interaction and a Leu-Leu core-core *a-a'* interaction. Following prior work in the Vinson laboratory,[9] we constructed A-BD_{CC}^{245} (sequence in Table 2.1). The modification added 17 residues at the N-terminus and replaced 6 out of 9 of the most N-terminal residues of the designed region (Table 2.1). Interestingly, A-BD_{CC}^{245} showed much greater helicity than BD_{CC}^{245} and BD_{CC}^{231}, indicating that either some of the N-terminal 26 residues and/or the distal C-terminal region are likely helical in this context (Fig. 2.3f). The T_m for A-BD_{CC}^{245} was similar to those for BD_{CC}^{231} and BD_{CC}^{245} (Table 2.1), whereas interaction with B-$BZLF1^{245}$ was significantly stabilized compared to the BD_{CC}^{245}/B-$BZLF1^{245}$ interaction as expected (Fig. 2.3f). The heterocomplex melted at > 80 °C (Table 2.2). Together these observations indicate that changes made in A-BD_{CC}^{245} did not stabilize the design homodimer, but further enhanced its interaction with B-$BZLF1^{245}$, as desired for inhibitor design.

For comparison, we constructed several other peptides with acidic extensions and assessed their self-association (Table 2.1). This modification dramatically destabilized $BZLF1^{245}$ by 28 °C (71 °C for $BZLF1^{245}$ vs. 43°C for A-$BZLF1^{245}$). A-$BZLF1^{231}$ was also destabilized, but by only 10 °C (43°C for $BZLF1^{231}$ vs. 33 °C for A-$BZLF1^{231}$). BD_{IED}^{245} was destabilized by an amount that could not be quantified because A-BD_{IED}^{245} did not exhibit a cooperative melt. A-$BZLF1^{245}$ was tested for interaction with B-$BZLF1^{245}$ and formed a heterocomplex with T_m of 74 °C (compared

to the T_m for B-BZLF1^{245} self-interaction, 67 °C, Tables 2.1, 2.2). The T_m for the heterocomplex between A-BZLF1^{231} and B-BZLF1^{231} was 58 °C (compared to the T_m for B-BZLF1^{231} self-interaction, 31 °C). These results are consistent with an interaction between the acidic extension and the basic region stabilizing the heterocomplexes, and also with an unfavorable interaction between the distal CT and the acidic extension, which is considered further in the Discussion.

Inhibiting DNA binding by BZLF1

We used an electrophoretic mobility shift assay (EMSA) to assess inhibition of B-BZLF1^{245} binding to DNA by different designed peptides (Fig. 2.6). The dimerization domain of BZLF1 lacking the basic region, BZLF1^{245}, was included for comparison purposes. All peptides tested showed concentration-dependent inhibition. BD^{245}_{CC}, A-BZLF1^{245} and A-BD^{245}_{CC} were more effective than BZLF1^{245}. Design BD^{245}_{IED} was also an effective inhibitor. The most potent inhibitor was A-BD^{245}_{CC}, which completely inhibited B-BZLF1^{245} binding to DNA at equi-molar concentration.

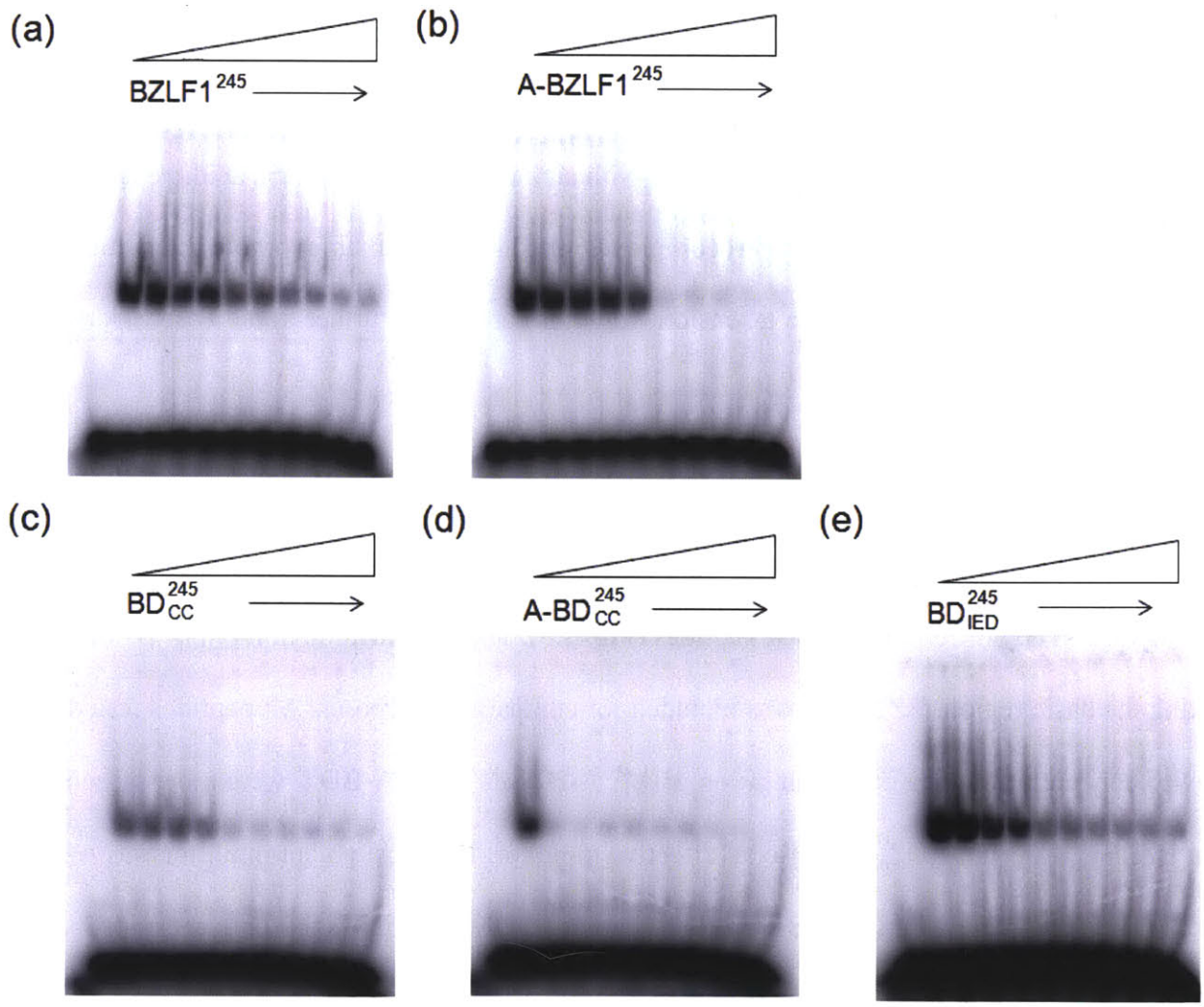


Figure 2.6 Peptide inhibition of B-BZLF1^{245} binding to DNA.
Representative gel-shift images were shown for: (a) BZLF1^{245}, (b) A-BZLF1^{245}, (c) BD^{245}_{CC}, (d) A-BD^{245}_{CC}, (e) BD^{245}_{IED}. The first two lanes for each gel include DNA only (first lane) and B-BZLF1^{245} with DNA (second lane). Inhibitor peptides were added in increasing concentrations from 10 nM to above 2 μM (left to right, 2-fold dilutions). Conditions are described in Materials and Methods in more detail, and were slightly different for panel (a)-(d) vs. panel (e).

Discussion

In this study, we employed different design strategies to create inhibitor peptides targeting the viral bZIP protein BZLF1. We sought peptides that achieved hetero-specificity through enhanced affinity for the target and/or reduced self-interaction. Below we discuss our different design

approaches and the experimental behaviors of our designed peptides.

Applying CLASSY to BZLF1

As demonstrated earlier,[18] CLASSY is an algorithm that can be applied to design bZIP-like coiled coils. It was developed in conjunction with a specialized scoring function that includes computed structure-based terms, helix propensities, and experimentally determined coupling energies. The scoring function was validated on a large-scale dataset of human bZIP coiled-coil interactions[13] and supported the successful design of numerous bZIP-binding peptides. It is not known to what extent the bZIP scoring function can be applied in design problems involving coiled-coil targets with features not observed in typical human bZIPs. Here, we explored whether the BZLF1 dimerization domain could be treated as a standard bZIP target for CLASSY design.

To treat BZLF1 as a coiled coil, we designed against the N-terminal part of the sequence and did experimental tests using constructs that did not include the distal CT (the "231" constructs, Fig 1b, 2a), much of which is not observed in the X-ray structure. The BZLF1 coiled-coil region is rather short (4 heptads), has only one Leu at position *d* among these heptads, and includes a region with very narrow inter-helical distance (~4 Å Cα-Cα distance at *a*-position residue 204). These variations might be expected to compromise performance of the scoring function, as coiled-coil context is known to influence the contributions of residues and residue pairs to stability.[17, 37, 38] Thus, methods validated using human bZIPs might not generalize broadly to all coiled-coil dimers. However, we found that design BD_{CC}^{231} incorporated elements very commonly employed in published anti-human bZIP designs (see below), and that these gave good experimental performance in this less canonical example. Success might be attributed to the fact that introducing more canonical residues at interfacial sites on one helix (the design) makes the

design-target heterodimer more similar to the human bZIPs, e.g. the heterodimer likely has a more typical helix-helix separation.

Features contributing to the stability and specificity of the designs

Analysis of the designed sequences suggests that stability and specificity were achieved using different combinations of core, edge and core-edge interactions. For example, in the BD_{CC}^{231} design, the *a* and *d* heptad positions were populated with hydrophobic Ile and Leu, respectively, (e.g. Y200L, A204I, K207L), which are expected to be exceptionally stabilizing in the design homodimer.[39] Therefore, a strategy that used only these mutations to stabilize the design-target interaction would likely stabilize the design self-interaction even more, and fail to achieve heterospecificity. Negative design elements that likely compensate for over-stabilization of the design self-interaction come from interfacial *e* and *g* positions occupied by negatively charged amino acids. These negative charges make favorable interactions with positively charged residues in the target (e.g. 201R, 207K), consistent with improving the stability of the design-target interaction. However, they also introduce repulsive *g-e'* or *e-g'* interactions in the design homodimer (e.g. 196E-201E (*g-e'*), 203E-208E (*g-e'*), 201E-203E (*e-g'*)). Similar examples of using a highly hydrophobic core to achieve stability while modulating specificity using interfacial charge have been observed in many prior coiled-coil designs.[32] One less familiar feature in the BD_{CC}^{231} design is the presence of an N-terminal glutamate at a *d* position. Two glutamate residues at *d* and *d'* in a homodimer are destabilizing in coiled coils,[40] but this residue potentially interacts favorably with an *e'* lysine in BZLF1, via a core-to-edge type interaction that has previously been noted in CLASSY-derived designs and other studies.[17, 18, 41, 42, 43]

Designs BD^{231}_{IED} and BD^{231}_{EEI} relied much more on core-to-edge interactions, which were placed close to the middle of the coiled coil in these designs. In contrast to *g-e'* interactions, no coupling energies have been measured for negatively charged residues at *d-d'* or *d-e'* sites. CLASSY performed poorly in predicting the relative stabilities of complexes involving BD^{231}_{IED} and BD^{231}_{EEI}, most likely because experimental data describing such charged core-core and core-edge interactions were not available to guide the development of the scoring function.[13,18] Nevertheless, a cluster of 4 negatively charged residues in the design homodimer proved very effective as a negative design element; BD^{231}_{IED} and BD^{231}_{EEI} did not appreciably self-associate. Affinity for the target was also compromised, however. Substitution of alanine with isoleucine at *a* position 204 was introduced to compensate for some of the lost stability of the heterodimer, showing how a different combination of stabilizing and destabilizing elements can generate a hetero-specific design that inhibits DNA binding (Fig. 2.6).

Substitution of isoleucine for alanine at *a* position 204 is found in all 3 designs. In the native structure, alanine at this position fits well in the tight space between unusually close helices (~4 Å Cα-Cα distance between residue 204 on the two chains). Isoleucine cannot be built into this site in the crystal structure without severe clashes. Nonetheless, the larger Ile was accommodated in all three designs, and an alanine to isoleucine mutation is stabilizing in the context of $BZLF1^{245}$ (an increase of T_m by 9 °C under the conditions of Table 2.1, data not shown). These data suggest a change in the backbone structure upon making this substitution. Local rearrangement of the design-BZLF1 complex to a more typical backbone structure probably helps explain why the CLASSY bZIP scoring function worked well. To achieve good predictive ability for a wider range of backbone structures, backbone flexibility could be treated explicitly.[43, 44]

The influence of the distal CT region

Previous studies revealed that the distal CT, although unresolved in the BZLF1 crystal structure, might interact with the N-terminal part of the BZLF1 coiled-coil region, thereby stabilizing the dimer.[27] We confirmed a stabilizing role for this region (Table 2.1, comparing $BZLF1^{231}$ and $BZLF1^{245}$). Interestingly, this effect depends on the sequence in the coiled-coil region (Table 2.1, 2.2). The distal CT does not stabilize the BD_{CC}^{245} design self-interaction, and it enhances the stability of the BD_{CC}^{245}-target interaction only modestly. On the other hand, the distal CT significantly increased the stability of the BD_{IED}^{245} design self-interaction, as well as the stability of the BD_{IED}^{245}-target interaction. There are more sequence changes in the BD_{CC} design, and the number of negative charges introduced is larger than in the BD_{IED} design. As discussed below, the influence of the distal CT is also sensitive to the acidic extension included in some designs. Although the structure of the interaction between the distal CT and the N-terminal part of the coiled coil in the native protein is not known, repulsive electrostatics, or unfavorable desolvation of charges in the coiled-coil region are plausible mechanisms for disfavoring this interaction in the BD_{CC} design.

Specificity against human bZIPs

We did not consider specificity against human bZIPs in our design procedure. However, we showed that the design BD_{CC} is not promiscuous in binding human bZIP proteins. Computational analysis predicted that the coiled-coil region of BD_{CC} would interact with the BZLF1 coiled coil moderately more favorably than with any other human bZIP coiled coil (but with a few close competitors). This is interesting, given the fairly canonical coiled-coil sequence features of BD_{CC}. The requirement to satisfy hydrogen bonding for Asn 204 at the *a* position in BD_{CC}, and the

charge complementarity between the *e* and *g* positions of BD_{cc} and BZLF1 helices but not most human proteins, contributed to the predicted binding preference.

Thermal stability studies confirmed that BD_{cc}^{245} does not bind strongly to selected human bZIPs identified in the computational analysis. In addition to selectivity derived from the coiled-coil region (which was predicted to be modest), the CT region likely confers additional specificity. Interactions with BD_{cc}^{245} and B-BZLF1^{245} could benefit from native-like contacts between the CT region and the coiled coil domain, which are not conserved in complexes with human proteins. Thus, the interaction specificity of BD_{cc}^{245} is likely encoded in both its coiled-coil domain and the CT region.

Improving inhibitor potency using an N-terminal acidic extension

The Vinson group has demonstrated that dominant-negative inhibitors of bZIP dimerization and DNA binding can be created by replacing the basic region of native or modified native bZIPs with an acidic sequence.[7] In this study, we used this strategy to improve the potency of our designed peptides. The resulting A-BD_{cc}^{245} peptide maintained specificity, showing little change in the T_m for the design self-association. The small change in homodimer stability probably results from destabilization by the negative charges in the extension, countered by a stabilizing leucine residue introduced at *d* position 193 (this residue is Glu in BD_{cc}).[9] A-BD_{cc}^{245} formed a more stable complex with the target B-BZLF1^{245} than did BD_{cc}^{245} (an increase of T_m > 14 °C at 4 μM, Table 2.2). This indicates that the acidic extension, which targets the basic region of bZIPs, can be used in conjunction with computational design methods targeting the coiled coil. Given that the Vinson laboratory has demonstrated that the coiled-coil region of A-ZIPs governs interaction specificity, while the acidic extension provides much enhanced affinity, this is an appealing

strategy for expanding the design of tight-binding and selective bZIP inhibitors.[7, 8, 9, 10, 18]

Interestingly, modifying BZLF1 with an acidic extension did not stabilize interaction of A-BZLF1^{245} with B-BZLF1^{245} as much as expected (T_m of 74 °C compared to 67 °C for the B-BZLF1^{245} homodimer, Table 2.1, 2.2). In contrast, interaction of the shorter construct A-BZLF1^{231} with B-BZLF1^{231} was stabilized to a much greater extent (T_m of 58 °C compared to 31 °C for the B-BZLF1^{231} homodimer). Furthermore, the destabilizing effect of the acidic extension on design homodimer stability is quite different in BZLF1^{245} vs. BZLF1^{231} (decreasing T_m values by 28 °C vs. 10 °C, Table 2.1). These observations are consistent with a model where the distal CT interacts unfavorably with the acid extension, much as it appears to interact unfavorably with negative charges in the N-terminal part of the BD_{cc} design. Although not addressed in the present study, the performance of A-BZLF1^{245} as an inhibitor could potentially be improved by redesigning the acidic extension so that interference from the distal CT is minimized, although this is difficult in the absence of structural information about this part of the protein.

Analysis of inhibitor potency

To test the designed peptides as inhibitors of BZLF1 DNA binding, we used an *in vitro* EMSA assay to monitor the population of B-BZLF1^{245} bound to DNA in the presence of different peptides (Fig. 2.6). It is unsurprising that A-BD_{cc}^{245}, which formed the most thermo-stable complex with B-BZLF1^{245} and exhibited the largest difference in homodimer vs. heterodimer stability, was the most potent inhibitor. The improved performance of BD_{cc}^{245} and A-BZLF1^{245} relative to the native peptide, BZLF1^{245}, could be rationalized by their improved affinity and/or anti-homodimer specificity (see below). BD_{IED}^{245} inhibited DNA binding effectively and we estimate its potency is similar to that of BZLF1^{245}, although these two peptides could not be compared using identical

assay conditions (see Materials and Methods). The effectiveness of BD_{IED}^{245} resulted from a combination of reduced affinity but improved anti-homodimer specificity.

To explore more generally how affinity and specificity each influence potency, we constructed a simple computational model with the following assumptions: 1) the target bZIP, the DNA, and the designed peptide were the only components present, 2) the target bZIP homodimer was the only species that could bind DNA (i.e. complete cooperative binding), 3) non-specific DNA binding was neglected. Some of the assumptions made may not apply to all of our experiments. We computed concentration dependent inhibition of DNA binding for a series of designs covering a spectrum of affinities and specificities. Affinity was described by the ratio between the dissociation constant of the target bZIP homodimer and that of the design-target heterodimer ($K_d^{T_2}/ K_d^{DT}$, D: design, T: target, see Materials and Methods), and specificity was described by the ratio between the dissociation constant for the design homodimer and that of the design-target heterodimer ($K_d^{D_2} / K_d^{DT}$). The efficacy of different inhibitors is illustrated in a heat map in Fig. 2.7 that indicates the improvement in IC_{50} over a reference for which $K_d^{D_2} = K_d^{DT} = K_d^{T_2}$. The reference inhibitor with affinity and specificity of 1 was included to reflect the behavior of the dimerization domain of the target bZIP. We explored two scenarios that led to different inhibition landscapes: one where modeled dissociation constants for the target bZIP complex and bZIP-DNA interactions were lower than the target bZIP concentration (Fig. 2.7a), and another where they were higher (Fig.2.7b)

The results in Fig. 2.7 support intuition about the importance of both affinity and specificity. Lines of constant color running across the plots in Fig. 2.7 show that equivalent potency can be achieved using different combinations of affinity and specificity. Clearly, neither affinity nor preference for hetero vs. homodimerization correlates directly with design performance. For the

purposes of discussion, we label 3 regions on the plots: $H_{affinity}$:L_{spec} indicates inhibitors with high affinity for the target but limited anti-homodimer specificity, $L_{affinity}$:H_{spec} indicates inhibitors with affinity for the target that is comparable to or weaker than the reference inhibitor, but with weaker self-association, and $H_{affinity}$:H_{spec} inhibitors have both tighter target-binding affinity and weaker self-association than the reference. Among our designs, and to the extent that approximate stabilities assessed by thermal denaturation under CD conditions can be extrapolated to the gel-shift assay, BD_{CC}^{245} and BD_{IED}^{245} are both $L_{affinity}$:H_{spec} inhibitors that use anti-homodimer specificity to improve inhibitor potency. A-BD_{CC}^{245} maintains anti-homodimer specificity but gains additional affinity via the acidic extension, making it a $H_{affinity}$:H_{spec} inhibitor.

The model in Fig. 2.7 is useful for broadly guiding the computational design of specific inhibitors, so we conclude with a few general observations. First, heterospecificity is important, but not sufficient, for good performance. A design is hetero-specific if the ratio $K_d^{T_2} \cdot K_d^{D_2}/(K_d^{DT})^2$ is larger than 1. In the figure, this region is below the dashed line and all inhibitors with potency better than the reference lie in this region. Maintaining hetero-specificity for high affinity designs imposes a bound on design homodimer stability. This is relevant for parallel dimeric coiled-coil targets, because amino-acid changes that enhance interaction with the target often stabilize the design self-interaction even more.[39] Second, the relative importance of improving affinity vs. specificity depends on the target and assay conditions. For panel a, improved hetero-specificity implies enhanced design performance regardless of whether affinity or specificity is the main contributor. On the other hand, if the target bZIP concentration is lower, as in panel b, improving specificity alone is no longer sufficient, and affinity must be optimized; very potent designs in panel b can only be achieved by optimizing along the path toward $H_{affinity}H_{spec}$. Finally, the overall diagonal trends for constant-IC_{50} regions in both panels emphasize that improving either

affinity or specificity can potentially lead to success, depending on the specific conditions and requirements for an application. Designs belonging to the $H_{affinity}H_{spec}$ class are the most effective. However, such designs might not exist, or could be hard to identify for a particular problem. In such cases, one could consider optimizing primarily affinity or specificity, depending on which is easier to achieve. Although not used extensively for this purpose here, the CLASSY algorithm is well suited for identifying designs with different affinity vs. specificity trade-offs.[18]

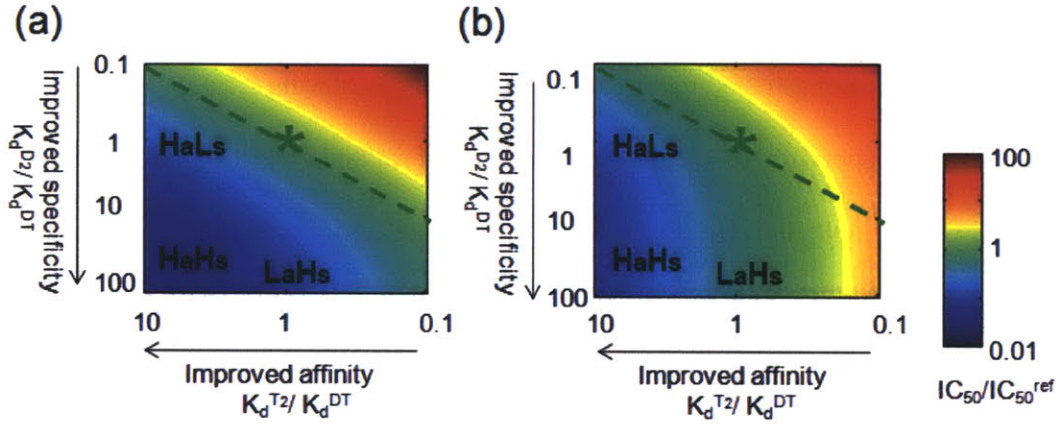


Figure 2.7 Inhibition of DNA binding as a function of the affinity and anti-homodimer specificity of the inhibitor.
A description of the model is given in Methods. The ratio of the IC_{50} for a design to the IC_{50} for a reference inhibitor with affinity equal to the wild-type protein is used as an indicator of design potency (scale at right). This ratio is plotted as a function of the affinity and specificity of the inhibitor. In (a), the K_d values for target dimerization and DNA binding are 10-fold lower than the bZIP concentration. In (b) the K_d values for both associations are 10-fold higher than the bZIP concentration. Labeling on the graph (H_aL_s: $H_{affinity}L_{spec}$, L_aH_s: $L_{affinity}H_{spec}$ and H_aH_s: $H_{affinity}L_{spec}$) is described in Discussion. The dashed line represents designs with zero hetero-specificity. The reference inhibitor is indicated with a star.

Conclusion: implications for protein design

This study addresses three topics relevant to the design of peptides that inhibit native protein-protein interactions. First is the issue of specificity, which arises in many protein design problems and is acute for coiled-coil targets where self-association of the design can compete with target inhibition. Using BZLF1 as a target, we characterized peptides that balance affinity

and specificity in different ways. This adds to the small number of examples where affinity and specificity have both been treated as design considerations.[18, 41, 43, 45, 46, 47, 48, 49] Second, we explored a design problem where features of the target that are not well described in an existing structure (the BZLF1 distal CT) nevertheless influence complex stability. We showed that different designs responded differently to the introduction of the distal CT. This argues for developing methods that broadly survey design solution space and discovering a large set of potentially good designs, rather than identifying only "the best" design according to some imperfect criteria. This can be accomplished in various ways, e.g. by exploring a range of tradeoffs between stability and specificity, or exploring a variety of related structural templates as design scaffolds.[18, 50] Testing diverse solutions maximizes the chance of finding a design that interacts well with poorly characterized features of the target. Finally, our best design exploited a modular strategy where optimization of the coiled-coil dimerization interface was coupled with a more generic strategy developed previously for stabilizing inhibitor-bZIP complexes. Modularity is likely to be a key strategy for the design of ever more complex molecular parts.

Materials and Methods

Cloning, protein expression and purification

Synthetic genes encoding native or redesigned BZLF1 sequence, residues 175 or 191 to 245 (B-BZLF1^{245}, BZLF1^{245}, BD$^{245}_{CC}$, BD$^{245}_{IED}$), were constructed by gene synthesis. Primers were designed using DNAWorks,[51] and a two-step PCR procedure was used for annealing and amplification. Genes encoding the native or redesigned sequence in the context of residues 191

to 231 were made in a single-step PCR reaction using the longer constructs as templates. The genes were cloned via BamHI/XhoI restriction sites into a modified version of a pDEST17 vector that encodes an N-terminal 6xHis tag and a GESKEYKKGSGS linker that improves the solubility of the recombinant protein.[28] To facilitate cloning of genes encoding the acidic extension, a pET16b vector (Novagen) was modified to encode an N-terminal 6xHis tag, followed by a GSY linker and the acidic extension sequence. Genes encoding $BZLF1^{231}$, $BZLF1^{245}$ and the designs BD^{245}_{CC} and BD^{245}_{IED} were subsequently cloned into the modified vector using AflII/XhoI restriction sites to make A-$BZLF1^{231}$, A-$BZLF1^{245}$, A-BD^{245}_{CC} and A-BD^{245}_{IED}. Recombinant proteins were expressed in *E. coli* RP3098 cells. Cultures were grown at 37 °C to an OD of ~0.4-0.9, and expression was induced by addition of 1 mM IPTG. Purification was performed under denaturing conditions (6M GdnHCl) using an Ni-NTA affinity column followed by reverse-phase HPLC. Human bZIP constructs containing the basic region and the coiled-coil domain were described previously.[28]

Computational protein design using CLASSY

The sequence BD_{cc} was designed using the CLASSY algorithm as previously reported.[18] In brief, the algorithm solves for the sequence predicted to interact most favorably with a target sequence (here, chosen to be the N-terminal part of the BZLF1 leucine zipper, residues 191 to 209) using integer linear programming. It is possible to impose constraints on the gap between the energy of interaction with the target and the energy of undesired states such as the design homodimer. No such constraint was applied in the design of BD_{cc}, which was predicted to favor the design-target interaction over design homodimerization without it. The scoring function used

was HP/S/Cv. This function was derived by combining molecular mechanics calculations and experimentally determined coupling energies for many core *a-a'* interactions.[13, 16] The Leu-Leu core *d-d'* interaction was modeled with an empirical value of –2 kcal/mol^{-1}. The HP/S/Cv structure-based energy function was transformed into a sequence-based expression using cluster expansion, and modified using empirical data, as described by Grigoryan et al.[18]

Predicting interactions between BD_{cc} and human bZIPs

BZLF1 was aligned with 36 human bZIPs using the conserved basic region, and interaction scores for residues 191-221 of BD_{cc} with the correspondingly aligned 31 residues of each human bZIP were computed using the HP/S/Cv model as described above.

Circular dichroism spectroscopy

Circular dichroism experiments were performed and analyzed, and T_m values fitted as described previously.[18] Thermal melts from 0 °C to 85 °C were mostly reversible, regaining ≥95% of signal or giving closely similar T_m values for the reverse melt (except for samples containing NFIL3, which precipitated upon heating to 85 °C). Melting temperatures were estimated by fitting the data to a two-state equilibrium (unfolded/folded), assuming no heat capacity changes upon folding. A detailed description of the equation was described previously.[18] In cases where high-temperature unfolding precluded accurate fitting of unfolded baselines, the T_m was either defined as the mid-point of the unfolding transition after manually picking the baseline (for the 1:1 mixture of B-BZLF1^{245} and A-BZLF1^{245}), or a lower bound on the T_m value

was estimated (for the 1:1 mixture of B-BZLF1^{245} and A-BD$^{245}_{CC}$). The protein concentrations are given in the figure legends. All measurements were performed in PBS buffer containing 12.5 mM potassium phosphate (pH 7.4), 150 mM KCl, 0.25 mM EDTA and 1 mM DTT. Samples were heated to 65 °C for 5 minutes before measurement to equilibrate peptide mixtures, and then cooled to and equilibrated at the starting temperature.

Analytical ultracentrifugation

Protein samples were dialyzed against the reference buffer (12.5 mM sodium phosphate, 150 mM NaCl, 1mM DTT, 0.25 mM EDTA, pH 7.4) three times (including once overnight) before measurements. Sedimentation equilibrium runs were performed with a Beckman XL-I analytical ultracentrifuge using interference optics. Two concentrations for each protein sample were prepared (50 and 100 μM), and runs at 3 different speeds (28,000, 35,000 and 48,000 rpm) were carried out at 20 °C. Each run was ~ 20 h, and equilibrium was confirmed by negligible differences between the sample distribution in the cell over sequential scans. Data were analyzed globally with the program HeteroAnalysis[52], using a calculated[53] partial specific volume of 0.7275 ml/g (for the BD$^{231}_{CC}$/BZLF1^{231} mixture) or 0.7245 ml/g (for BD$^{231}_{CC}$) and a solution density of 1.005 g/ml.

Electrophoretic mobility shift assay (EMSA)

Gel shift assays were performed as described previously[28]. Briefly, 10 nM B-BZLF1^{245} was prepared either alone or mixed with each inhibitor at 9 concentrations ranging from 10 nM to 2560 nM in 2-fold dilutions. Gel-shift buffer ((150 mM KCl, 25 mM TRIS pH 8.0, 0.5 mM EDTA, 2.5 mM DTT, 1 mg/ml BSA, 10% (v/v) glycerol, 0.1 μg/ml competitor DNA (Poly

(I)·Poly (C) (Sigma))) was then added and incubated for 10 minutes at 42 °C. Closely similar results were obtained when incubating samples for 20 minutes at 42 °C. The competitor BD_{IED}^{245} was not stable upon heating and was incubated for 2 hours at 18-22 °C. Radiolabeled annealed AP-1 site ,CGCTTGATGACTCAGCCGGAA (IDT), at a final concentration of 0.7 nM was added and incubated for 15 minutes at 18-22 °C. Complexes were separated on NOVEX DNA retardation gels (Invitrogen). Dried gels were imaged using a phosphorimaging screen and a Typhoon 9400 imager. ImageQuant software (Amersham Biosciences) was used to quantify band intensities.

Simulating the impact of affinity and specificity on designed peptide behaviors

The simulation treated the following species: The target bZIP monomer (T), the target bZIP homodimer (T_2), the design monomer (D), the design homodimer (D_2), the design-target bZIP heterodimer (DT), free DNA (DNA) and the complex formed between the target bZIP homodimer and DNA (T_2DNA). Species are linked by the following reactions:

$$2T \leftrightharpoons T_2 \qquad \frac{[T]^2}{[T_2]} = K_d^{T_2}$$

$$2D \leftrightharpoons D_2 \qquad \frac{[D]^2}{[D_2]} = K_d^{D_2}$$

$$D + T \leftrightharpoons DT \qquad \frac{[D][T]}{[DT]} = K_d^{DT}$$

$$T_2 + DNA \leftrightharpoons T_2DNA \qquad \frac{[T_2][DNA]}{[T_2DNA]} = K_d^{T_2DNA}$$

$$[T] + [DT] + 2[T_2] + 2[T_2DNA] = [T]_{total}$$

$$[D] + [DT] + 2[D_2] = [D]_{total}$$

$$[DNA] + [T_2DNA] = [DNA]_{total}$$

Affinity is defined as $K_d^{T_2} / K_d^{DT}$, and a value > 1 indicates the design-target bZIP heterodimer is more stable than the target bZIP homodimer (improved affinity). Specificity is defined as $K_d^{D_2} / K_d^{DT}$, and a value > 1 indicates the design-target bZIP heterodimer is more stable than design homodimer (improved specificity). A design with affinity and specificity equal to 1 was used as a reference. The IC_{50} value was defined as the design concentration $[D]_{total}$ at which 50% less DNA is bound relative to zero design concentration. The total target bZIP concentration $[T]_{total}$ was fixed at 10 nM, and the total DNA concentration $[DNA]_{total}$ at 0.7 nM. Different combinations of $K_d^{T_2}$ and $K_d^{T_2DNA}$ values were explored (10^{-9}, 10^{-8}, and 10^{-7} M for each), including when both are lower than $[T]_{total}$ (10^{-9} M/10^{-9} M, Fig. 2.7a) and when both are higher than $[T]_{total}$ (10^{-7} M/10^{-7} M, Fig. 2.7b). For each combination of fixed $K_d^{T_2}$ and $K_d^{T_2DNA}$, the IC_{50} values for a range of designs with different affinities (0.1 to 10) and specificities (0.1 to 100) were calculated. The ratio $IC_{50}^{design}/IC_{50}^{ref}$, with a value < 1 implying greater potency than the reference, was plotted as a heat map. The dashed lines on the plots in Fig. 2.7 indicate points where the product of affinity and specificity $((K_d^{T_2} * K_d^{D_2})/(K_d^{DT} * K_d^{DT}))$ equals 1. All designs below the dashed line are hetero-specific. The simulation was carried out and heat maps were generated using Matlab (MathWorks).

Acknowledgements

We thank K. E. Thompson for designing the acidic extension vector, making the A-BZLF1[245] construct, and providing valuable suggestions. We thank G. Grigoryan for assistance with the CLASSY algorithm, and members of the Keating lab, especially O. Ashenberg, C. Negron, S. Dutta, L. Reich, V. Potapov, K. Hauschild and J. DeBartolo for helpful discussion of the

manuscript. We thank D. Pheasant at the Biophysical Instrumentation Facility at MIT for assistance in analytical ultracentrifugation experiments. A.W.R. was supported by a Koch graduate fellowship. This work was funded by NIH award GM067681 and used computer resources provided by NSF award 0821391.

References

1. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**, 539-544.
2. Rishi, V., Potter, T., Laudeman, J., Reinhart, R., Silvers, T., Selby, M., Stevenson, T., Krosky, P., Stephen, A. G., Acharya, A., Moll, J., Oh, W. J., Scudiero, D., Shoemaker, R. H. & Vinson, C. (2005). A high-throughput fluorescence-anisotropy screen that identifies small molecule inhibitors of the DNA binding of B-ZIP transcription factors. *Anal. Biochem.* **340**, 259-271.
3. Rishi, V., Oh, W. J., Heyerdahl, S. L., Zhao, J., Scudiero, D., Shoemaker, R. H. & Vinson, C. (2010). 12 Arylstibonic acids that inhibit the DNA binding of five B-ZIP dimers. *J. Struct. Biol.* **170**, 216-225.
4. Mason, J. M., Schmitz, M. A., Műller, K. M. & Arndt, K. M. (2006). Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 8989-8994.
5. Mason, J. M., Műller, K. M. & Arndt, K. M. (2007). Positive aspects of negative design: simultaneous selection of specificity and interaction stability. *Biochemistry* **46**, 4804-4814.
6. Mason, J. M., Hagemann, U. B. & Arndt, K. M. (2009). Role of hydrophobic and electrostatic interactions in coiled coil stability and specificity. *Biochemistry* **48**, 10380-10388.
7. Acharya, A., Rishi, V., Moll, J. & Vinson, C. (2006). Experimental identification of homodimerizing B-ZIP families in Homo sapiens. *J. Struct. Biol.* **155**, 130-139.
8. Krylov, D., Olive, M. & Vinson, C. (1995). Extending dimerization interfaces: the bZIP basic region can form a coiled coil. *EMBO J.* **14**, 5329-5337.
9. Olive, M., Krylov, D., Echlin, D. R., Gardner, K., Taparowsky, E. & Vinson, C. (1997). A dominant negative to activation protein-1 (AP1) that abolishes DNA binding and inhibits oncogenesis. *J. Biol. Chem* **272**, 18586-18594.
10. Ahn, S., Olive, M., Aggarwal, S., Krylov, D., Ginty, D. D. & Vinson, C. (1998). A dominant-negative inhibitor of CREB reveals that it is a general mediator of stimulus-dependent transcription of c-fos. *Mol. Cell. Biol.* **18**, 967-977.
11. Gerdes, M. J., Myakishev, M., Frost, N. A., Rishi, V., Moitra, J., Acharya, A., Levy, M. R., Park, S. W., Glick, A., Yuspa, S. H. & Vinson, C. (2006). Activator protein-1 activity

regulates epithelial tumor cell identity. *Cancer Res.* **66**, 7578-7588.
12. Oh, W. J., Rishi, V., Orosz, A., Gerdes, M. J. & Vinson, C. (2007). Inhibition of CCAAT/enhancer binding protein family DNA binding in mouse epidermis prevents and regresses papillomas. *Cancer Res.* **67**, 1867-1876.
13. Grigoryan, G. & Keating, A. E. (2006). Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.* **355**, 1125-1142.
14. Fong, J. H., Keating, A. E. & Singh, M. (2004). Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol.* **5**, R11.
15. Krylov, D., Mikhailenko, I. & Vinson, C. (1994). A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. *EMBO J.* **13**, 2849-2861.
16. Acharya, A., Rishi, V. & Vinson, C. (2006). Stability of 100 homo and heterotypic coiled-coil a-a' pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry* **45**, 11324-11332.
17. Steinkruger, J. D., Woolfson, D. N. & Gellman, S. H. (2010). Side-chain pairing preferences in the parallel coiled-coil dimer motif: insight on ion pairing between core and flanking sites. *J. Am. Chem. Soc.* **132**, 7586-7588.
18. Grigoryan, G., Reinke, A. W. & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859-864.
19. Newman, J. R. & Keating, A. E. (2003). Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**, 2097-2101.
20. Vinson, C., Acharya, A. & Taparowsky, E. J. (2006). Deciphering B-ZIP transcription factor interactions in vitro and in vivo. *Biochim. Biophys. Acta.* **1759**, 4-12.
21. Countryman, J., Jenson, H., Seibl, R., Wolf, H. & Miller, G. (1987). Polymorphic proteins encoded within BZLF1 of defective and standard Epstein-Barr viruses disrupt latency. *J. Virol.* **61**, 3672-3679.
22. Schepers, A., Pich, D. & Hammerschmidt, W. (1993). A transcription factor with homology to the AP-1 family links RNA transcription and DNA replication in the lytic cycle of Epstein-Barr virus. *EMBO J.* **12**, 3921-3929.
23. Feederle, R., Kost, M., Baumann, M., Janz, A., Drouet, E., Hammerschmidt, W. & Delecluse, H. J. (2000). The Epstein-Barr virus lytic program is controlled by the co-operative functions of two transactivators. *EMBO J.* **19**, 3080-3089.
24. Liu, P. & Speck, S. H. (2003). Synergistic autoactivation of the Epstein-Barr virus immediate-early BRLF1 promoter by Rta and Zta. *Virulogy* **310**, 199-206.
25. Young, L. S. & Rickinson, A. B. (2004). Epstein-Barr virus: 40 years on. *Nat. Rev. Cancer* **4**, 757-768.
26. Petosa, C., Morand, P., Baudin, F., Moulin, M., Artero, J. B. & Műller, C. W. (2006). Structural basis of lytic cycle activation by the Epstein-Barr virus ZEBRA protein. *Mol. Cell* **21**, 565-572.
27. Schelcher, C., Al Mehairi, S., Verrall, E., Hope, Q., Flower, K. B., B. , Woolfson, D. N., West, M. J. & Sinclair, A. J. (2007). Atypical bZIP domain of viral transcription factor contributes to stability of dimer formation and transcriptional function. *J. Virol.* **81**, 7149-7155.
28. Reinke, A. W., Grigoryan, G. & Keating, A. E. (2010). Identification of bZIP interaction partners of viral proteins HBZ, MEQ, BZLF1, and K-bZIP using coiled-coil arrays. *Biochemistry* **49**, 1985-1997.

29. Hicks, M. R., Al-Mehairi, S. S. & Sinclair, A. J. (2003). The zipper region of Epstein-Barr virus bZIP transcription factor Zta is necessary but not sufficient to direct DNA binding. *J. Virol.* **77**, 8173-8177.
30. O'Shea, E. K., Lumb, K. J. & S., K. P. (1993). Peptide 'Velcro': design of a heterodimeric coiled coil. *Curr. Biol.* **3**, 658-667.
31. Vinson, C., Myakishev, M., Acharya, A., Mir, A. A., Moll, J. R. & Bonovich, M. (2002). Classification of human B-ZIP proteins based on dimerization properties. *Mol. Cell. Biol.* **22**, 6321-6335.
32. Woolfson, D. N. (2005). The design of coiled-coil structures and assemblies. *Adv. Protein Chem.* **70**, 79-112.
33. Grigoryan, G. & Keating, A. (2008). Structural specificity in coiled-coil interactions. *Curr. Opin. Struct. Biol.* **18**, 477-483.
34. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401-1407.
35. Taylor, C. M. & Keating, A. E. (2005). Orientation and oligomerization specificity of the Bcr coiled-coil oligomerization domain. *Biochemistry* **44**, 16246-16256.
36. Mason, J. M. & Arndt, K. M. (2004). Coiled coil domains: stability, specificity, and biological implications. *Chembiochem* **5**(170-176).
37. Moitra, J., Szila'k, L., Krylov, D. & Vinson, C. (1997). Leucine is the most stabilizing aliphatic amino acid in the d position of a dimeric leucine zipper coiled coil. *Biochemistry* **36**, 12567-12573.
38. Lu, S. M. & Hodges, R. S. (2004). Defining the minimum size of a hydrophobic cluster in two-stranded alpha-helical coiled-coils: effects on protein stability. *Protein Sci.* **13**, 714-726.
39. Acharya, A., Ruvinov, S. B., Gal, J., Moll, J. R. & Vinson, C. (2002). A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K. *Biochemistry* **41**, 14122-14131.
40. Tripet, B., Wagschal, K., Lavigne, P., Mant, C. T. & Hodges, R. S. (2000). Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 amino acid substitutions in position "d". *J. Mol. Biol.* **300**, 377-402.
41. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**, 45-52.
42. Reinke, A. W., Grant, R. A. & Keating, A. E. (2010). A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *J. Am. Chem. Soc.* **132**, 6025-6031.
43. Barth, P., Schoeffler, A. & Alber, T. (2008). Targeting metastable coiled-coil domains by computational design. *J. Am. Chem. Soc.* **130**, 12038-12044.
44. Mandell, D. J. & Kortemme, T. (2009). Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420-428.
45. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nat. Struct. Biol.* **11**, 371-379.
46. Ali, M. H., Taylor, C. M., Grigoryan, G., Allen, K. N., Imperiali, B. & Keating, A. E. (2005). Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed alpha/beta structure. *Structure* **13**, 225-234.
47. Bolon, D. N., Grant, R. A., Baker, T. A. & Sauer, R. T. (2005). Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12724-12729.

48. Sammond, D. W., Eletr, Z. M., Purbeck, C. & Kuhlman, B. (2010). Computational design of second-site suppressor mutations at protein-protein interfaces. *Proteins* **78**, 1055-1065.
49. Karanicolas, J. & Kuhlman, B. (2009). Computational design of affinity and specificity at protein-protein interfaces. *Curr. Opin. Struct. Biol.* **19**, 458-463.
50. Fu, X., Apgar, J. R. & Keating, A. E. (2007). Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-xL. *J. Mol. Biol.* **371**, 1099-1117.
51. Hoover, D. M. & Lubkowski, J. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43.
52. Cole, J. L. & Lary, J. W. (2006). Heteroanalysis, Analytical Ultracentrifugation Facility, University of Connecticut, Storrs, CT
53. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. (1992). Computer aided interpretation of analytical sedimentation data for proteins. *Analytical Ultracentrifugation in Biochemistry and Polymer Science. Cambridge, Royal Society of Chemistry*, 90-125.
54. Glover, J. N. & Harrison, S. C. (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**, 257-261.

Chapter 3

Design of anti-apoptotic Bcl-2 proteins with novel interaction specificity toward different BH3 peptides

This work is currently being prepared as a manuscript to be submitted later

Collaborator notes:

Hector Palacios cloned and purified some of the recombinant BH3 peptides and cloned the first designed library.

Introduction

Engineering protein-protein interactions is critical to numerous areas of basic science and biotechnology. Designed proteins can be used to inhibit or activate other proteins[1,2], to alter native proteins to study their functions[3,4], or to create novel interactions that rewire cell signaling in synthetic biological systems[5,6]. Many design applications demand that interactions be specific, i.e. that designed proteins interact only with target proteins and not with off-targets. This adds an additional layer of complexity to what is already a difficult design challenge.

Commercially useful interaction reagents and therapeutics are almost always identified by screening large combinatorial libraries, which can be an inefficient process due to the enormity of sequence space. Recent progress in computational design[7,8,9] offers great promise to accelerate the discovery of protein reagents with desired properties, by predicting good binders at the outset. But current methods are limited by available binding models. Contemporary physics based structural models enjoy the advantage of being general, but are computationally inefficient and not always accurate[10,11], while models that include statistical terms derived from known structures[12] can address some of these deficiencies but still do not provide high confidence in results. Methods relying on evolutionary sequence analysis[13] or machine learning[14] can only be applied for protein families that satisfy certain criteria (e.g. having a large number of protein sequences with known interaction properties).

Recently, several groups have proposed and tested the idea of computationally designing a library of protein sequences[15-23]. The rationale is that given the imperfect models used in protein design, assaying a large number of designed sequences simultaneously will increase the overall success rate. Previous studies have suggested several important factors to be considered for library design. These include whether the library can be cloned cost-effectively[19], whether the

library is adequately covered by the experimental screening platform[17,18,19], and the diversity of the library sequences[17,22]. Trade-offs exist when considering these factors[22]. For example, including more diverse sequences can lead to a library size larger than is practical to assemble or screen.

In this study we examined these issues in the context of designing protein-protein interaction specificity. Specificity design imposes multiple objectives, i.e. binding to one desired target but not to a related competitor. The demands on the accuracy of any binding model are high, and the risk of designs not working as expected is also high[24]. Here we present a novel library design strategy that emphasizes maintenance of a high degree of useful diversity, as predicted by an imperfect model. Our framework consists of two stages. In the first, desired sequence features are predicted using the structural modeling software Rosetta [12]. Desired sequence features are defined permissively, and emphasize the selection of residues predicted to maintain binding to a desired target. In the second stage, we apply the optimization algorithm integer linear programming (ILP) to design a combinatorial, degenerate codon based DNA library encoding the desired sequence features. Different constraints, such as an upper limit on library size, can be introduced easily into the ILP optimization, making it possible to rigorously explore the different trade-offs in library design as described above.

We applied this framework to the design of anti-apoptotic Bcl-2 proteins with novel peptide interaction specificity. The anti-apoptotic Bcl-2 family[25] proteins have a globular, helical fold and bind to short helical peptides derived from pro-apoptotic proteins, here called BH3 peptides. Native Bcl-2 family proteins bind BH3 peptides with a range of different specificities[26,27]. We aimed to re-design anti-apoptotic protein Bcl-xL so that it would lose the ability to strongly interact with BH3 peptide Bim but still bind tightly to a BH3 peptide derived from Bad. This is

an interesting specificity design problem because all known anti-apoptotic Bcl-2 proteins interact strongly with Bim, which is proposed as an "activator" BH3[28] in some models of the regulation of apoptosis[28,29,30]. In contrast, the BH3-only protein Bad, proposed as a "sensitizer", interacts with the anti-apoptotic proteins in a more selective manner. We designed libraries of Bcl-xL variants and screened these using yeast surface display[31]. We successfully obtained Bcl-xL variants that showed a strong preference for binding Bad over Bim. Detailed investigation of the sequence characteristics revealed that our inclusive design strategy was crucial for identifying high specificity sequences. We further showed that our designed protein is globally specific against binding 10 other BH3-only peptides not considered in library design, with interesting implications for specificity design involving multiple off-targets.

Results

Library design

There are two stages in our library design procedure (Fig. 3.1A). In the first stage, desired sequence features were identified. To simplify the analysis, we defined sequence features as individual residues, making the assumption that the energetic contribution of an amino acid at a given position is independent of its sequence context. This left the screening experiments to identify/avoid potential higher order interactions among residues at different designed positions, eliminating the possibility that such information could be predicted and used for making a more efficient library. The implications of this are considered further in the Discussion.

Guided by crystal structures of complexes between Bcl-xL and Bim/Bad[32,33], we chose 9 Bcl-xL sites where contacts are made to the central part of the Bim/Bad BH3 peptide for redesign. These 9 positions mostly interact with BH3 positions occupied by different amino acids in Bim vs. Bad (Fig. 3.1B, 3.1C). We next used the structural modeling suite Rosetta to predict how

different amino acids at each Bcl-xL designed position, in the sequence context of Bcl-xL, could affect interaction with Bim or Bad. Amino acids to be modeled at each position were chosen manually by considering factors such as hydrophobicity and size. Modeled complexes between different Bcl-xL point mutants and Bim or Bad were generated and their Rosetta energy scores relative to that of the native amino acid, ΔE_{Bim} and ΔE_{Bad}, were obtained (see Materials and Methods).

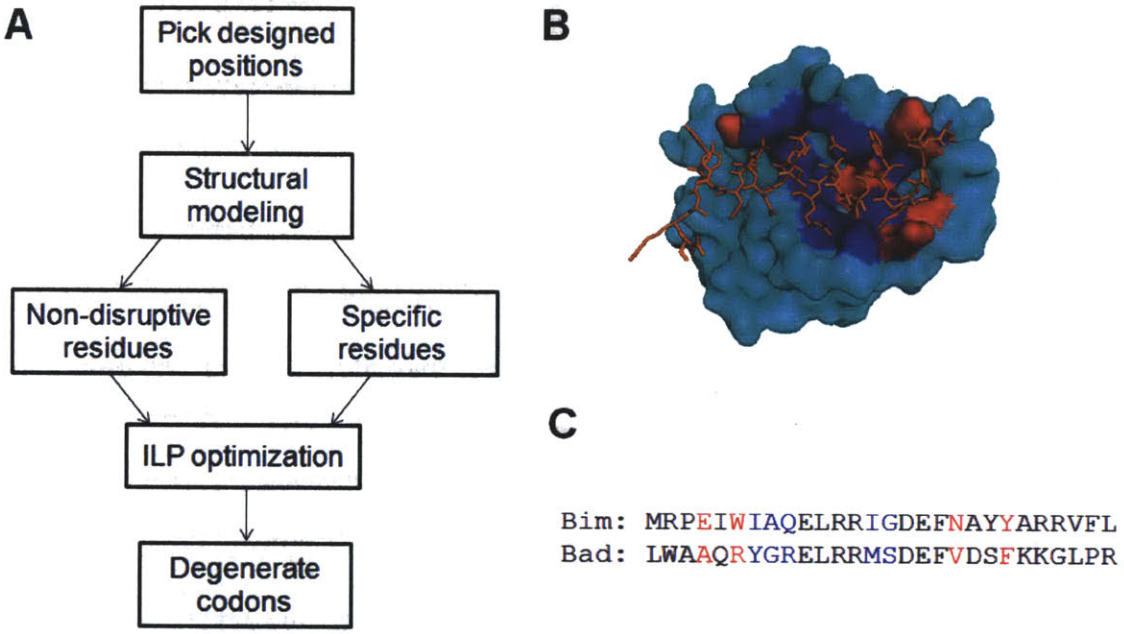


Figure 3.1 Library design protocol
(A) The library design protocol. Non-disruptive and specific residues were predicted at every designed position, and ILP optimization was performed to select degenerate codons that together maximize the inclusion of non-disruptive residues while enforcing the inclusion of specific residues, under a library size constraint. (B) The interface between Bcl-xL and a Bim BH3 peptide (PDB ID: 3FDL), with Bcl-xL shown in cyan and the peptide shown in orange sticks (PyMol, Delano Scientific). Designed positions for the first library are blue, while those for the second library are red. (C) A sequence alignment of the Bim (residue 142 to 169) and Bad (residue 104 to 131) BH3. Positions occupied by different amino acids and exploited for Bcl-xL design are colored in blue (for the first library) or red (for the second library).

Based on this analysis, we indentified non-disruptive residues that were predicted not to greatly disrupt binding to the desired target Bad. This is a less stringent and more inclusive prediction criterion than demanding that residues contribute to binding specificity. We argued that energies above a certain threshold could imply serious steric clashes or under-packing in the modeled complexes, and defined a ΔE_{Bad} value to serve as a cutoff for non-disruptive residues (Table 3.1). We also tabulated the difference in Rosetta energy scores for the modeled complexes between the corresponding Bcl-xL point mutant and Bim/Bad (ΔE_{Bim} - ΔE_{Bad}). Residues with a score difference above a certain threshold were predicted as specific residues that could contribute to the desired interaction specificity of Bad over Bim (Table 3.1). It should be noted that the specificity residues are a subset of the non-disruptive residues.

Table 3.1 The first designed library

Position	residues modeled[a]	residues encoded[b]
F97	AFGILMV	FILM (WTK)
Y101	AFGILMTVY	FHLY (YWT)
A104	AFGILMSTVY	ACDEFGIKLMNRSTVWYZ (DNK)
L108	AFGILMV	AGILPRSTV (VBT)
L112	AFGILMV	LMV (DTG)
V126	AFGILMV	AGIMRSTV (RBK)
E129	AEITV	AEIKTV (RHA)
L130	AFGILMV	ILV (VTC)
A142	AGSTV	AGST (RSC)

[a] Residues modeled at each position. Underlined residues were predicted to be non-disruptive, shaded residues were predicted to be specific.
[b] Residues included in the designed library (encoded by the degenerate codon in parentheses). A stop codon is indicated by "Z". The IUBMB abbreviations for mixture of nucleotides were adopted when representing the degenerate codons.

We proceeded to design a library, i.e. to optimize combinations of degenerate codons encoding diversity at each designed position. We formulated the optimization problem as an integer linear programming (ILP) problem. The objective to be maximized was the number of

unique library sequences with designed positions all occupied by non-disruptive residues. This is the product of the number of non-disruptive residues encoded by the degenerate codons across all designed position. Note that this objective could also be loosely interpreted as the number of unique protein sequences predicted to bind the desired target Bad. We enforced two constraints in the ILP problem. The first was on the library size in DNA space, which was set to 10^7, a conservative estimate for obtaining good sequence coverage in yeast surface display. The second was that all predicted specificity residues as well as all native residues were required to be included in the library. Both the objective and the size constraint are products and become linear in logarithm space, making the problem amenable to the ILP optimization (see Materials and Methods). The optimized library (Table 3.1) had a size of $8.9*10^6$ and contained $2.2*10^5$ unique protein sequences predicted to bind Bad, about ~6% of all library DNA sequences encode protein sequences predicted to bind Bad.

Yeast surface display screening

We used yeast surface display for the experimental screening. Native Bcl-xL displayed on the yeast surface bound both the Bim and Bad BH3 domains strongly, but did not bind the Noxa BH3 domain (data not shown), agreeing with previous binding studies done in solution. The designed library was enriched in sequences binding Bad, as expected. Approximately 5% of the population showed binding at 10 nM Bad BH3 (Fig. 3.2A). Interestingly, the designed library bound a Bim BH3 peptide even better than Bad (Fig. 3.2A), and this is discussed further below.

The designed library was subjected to 6 rounds of screening to identify Bcl-xL variants that bound Bad in preference to Bim (see Materials and Methods). The final population showed significantly enhanced specificity, with binding to Bad detectable at 1 nM Bad BH3, but that to Bim detectable only at 100 nM (Fig. 3.2B). Characterization of 48 clones randomly selected

from this population gave 21 unique clones with stronger binding signal under 1 nM Bad over 10 nM Bim (Fig. 3.2C, Table 3.2). The results revealed one Bcl-xL designed position, 142, at which substitution of Ala to Gly (A142G) was found in all sequences. Some designed positions were occupied by both native and non-native amino acids across all sequences, whereas some were occupied only with the native amino acid. A more detailed examination of the emerging sequence features and how they relate to the ones predicted in the library design is presented in the Discussion.

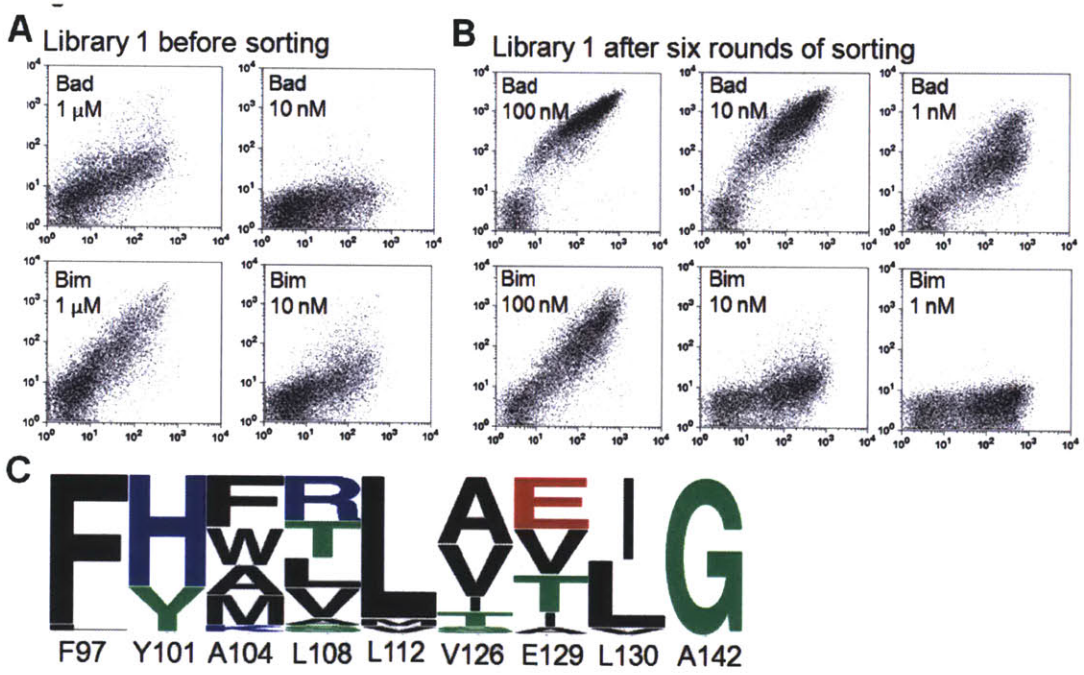


Figure 3.2 The first designed library
(A) Flow cytometry plots showing the first designed library displayed on the yeast surface binding Bad (top) and Bim (bottom) at 1 μM and 10 nM. (B) Flow cytometry plots showing the final sorted population from the first designed library displayed on the yeast surface binding Bad (top) and Bim (bottom) at 100 nM, 10 nM and 1 nM. (C) Sequence frequency plot for 21 unique sequences identified as specific for Bad over Bim BH3 from the first designed library (Table 3.2), with the native Bcl-xL residue shown below each column. Plots were generated using WebLogo[45].

Table 3.2 Unique sequences isolated from the first designed library

	F97	Y101	A104	L108	L112	V126	E129	L130	A142	Other[b]
A2	F	H[a]	W	R	L	V	I	I	G	
A3	F	Y	A	L	M	A	V	L	G	
A4	F	H	F	A	L	A	A	L	G	
B2	F	H	F	T	L	V	T	I	G	Q111K
B4	F	Y	A	V	V	A	E	V	G	
B5	F	Y	M	L	L	A	V	L	G	
C1	F	Y	M	V	L	A	E	I	G	
C3	F	H	W	S	L	V	T	I	G	
C4	F	H	F	T	L	V	V	I	G	
C5	F	H	K	L	L	A	I	L	G	S122I
C6	L	H	F	R	L	V	T	I	G	
D3	F	H	A	L	L	T	V	L	G	
D5	F	H	W	R	L	I	E	I	G	
D6	F	H	F	T	L	V	T	I	G	
E2	F	Y	M	V	L	S	E	I	G	
E5	F	H	W	R	L	I	T	I	G	
F4	F	Y	A	V	L	A	E	I	G	
G3	F	H	F	R	L	T	V	L	G	
G5	F	H	F	T	L	A	E	L	G	
H3	F	H	W	R	L	A	E	L	G	
H6	F	H	M	T	L	V	V	L	G	R103W

[a] Shaded residues were either not included in the modeling or not predicted as "non-disruptive" residues, but were included in the library due to degenerate codons.
[b] Residues under the "other" column were not included in the library design.

Design and screening of a second library with improved specificity

Based on the promising results of the library screen, we proceeded to design a second library to identify sequences with further improved specificity. Using the same structural modeling protocol described above, we predicted non-disruptive residues and specificity residues for 6 additional Bcl-xL positions (Fig. 3.1B, Table 3.3). These new positions were mostly located at the edge of the BH3-binding interface, and not surprisingly, our very relaxed definition of non-disruptive residues included almost all residues. Among the 9 designed positions screened in the previous library, we fixed position 142 as Gly (A142G), whereas position 97 and 112 were reverted back to the native residues. Non-disruptive residues at the other 6 positions (101, 104, 108, 126, 129, 130) were redefined as amino acids with significant frequency in the first round of screening (Fig. 3.2C, Table 3.2). A total of 12 Bcl-xL positions were randomized in the new library. The same ILP library optimization procedure described previously was carried out to select degenerate codons for these positions. To increase efficiency in encoding amino-acid diversity, we introduced a slight modification to allow some designed positions to be encoded by a pair of degenerate codons rather than just one, subject to constraints imposed by the PCR assembly protocol (See Materials and Methods).

Significant improvement in specificity was observed after two rounds of screening the newly designed library (Fig. 3.3A). Sequencing results revealed strong sequence bias at several designed positions (Fig. 3.3B, Table 3.4), as discussed below. We performed 5 additional rounds of screening, and the final population was highly specific for binding Bad over Bim (Fig. 3.3D), showing good binding to Bad at 1 nM Bad BH3 but much lower binding to Bim BH3 at 1 μM. Only 2 sequences were present in this population, L2-7-A1 and L2-7-F1 (Table 3.5). Each contained 9 mutations from native Bcl-xL, and the mutations were consistent with residues

observed at high frequency after two rounds of screening, as shown in Fig 3.3B. Interestingly, both sequences contained a mutation (F105L) not present in the designed library. The effect of this mutation was investigated and analyzed below.

Table 3.3 The second designed library[a]

	residues modeled	residues encoded
E96	ADEFGHIKLMNQRSTVY	EGLQRV (SDA)
Q111	ADEFGHIKLMNQRSTVY	ADEGHIKLMNPQRSTV (VNK)
S122	ADEFGHIKLMNQRSTVY	ADFHILNPSTVY (NHC)
Q125	ADEFGHIKLMNQRSTVY	ADEFGILNQRSTVY (DHT_SDA)
F146	AFGILMV	AFL (TTS_GCT)
Y195	FY	FY (TWC)
Y101	HY	HY (YAT)
A104	AFMW	AM (GCA_ATG)
L108	LRTV	LV (STG)
V126	AV	AV (GYA)
E129	ETV	EV (GWA)
L130	LI	LI (MTC)

[a] See descriptions for Table 3.1.

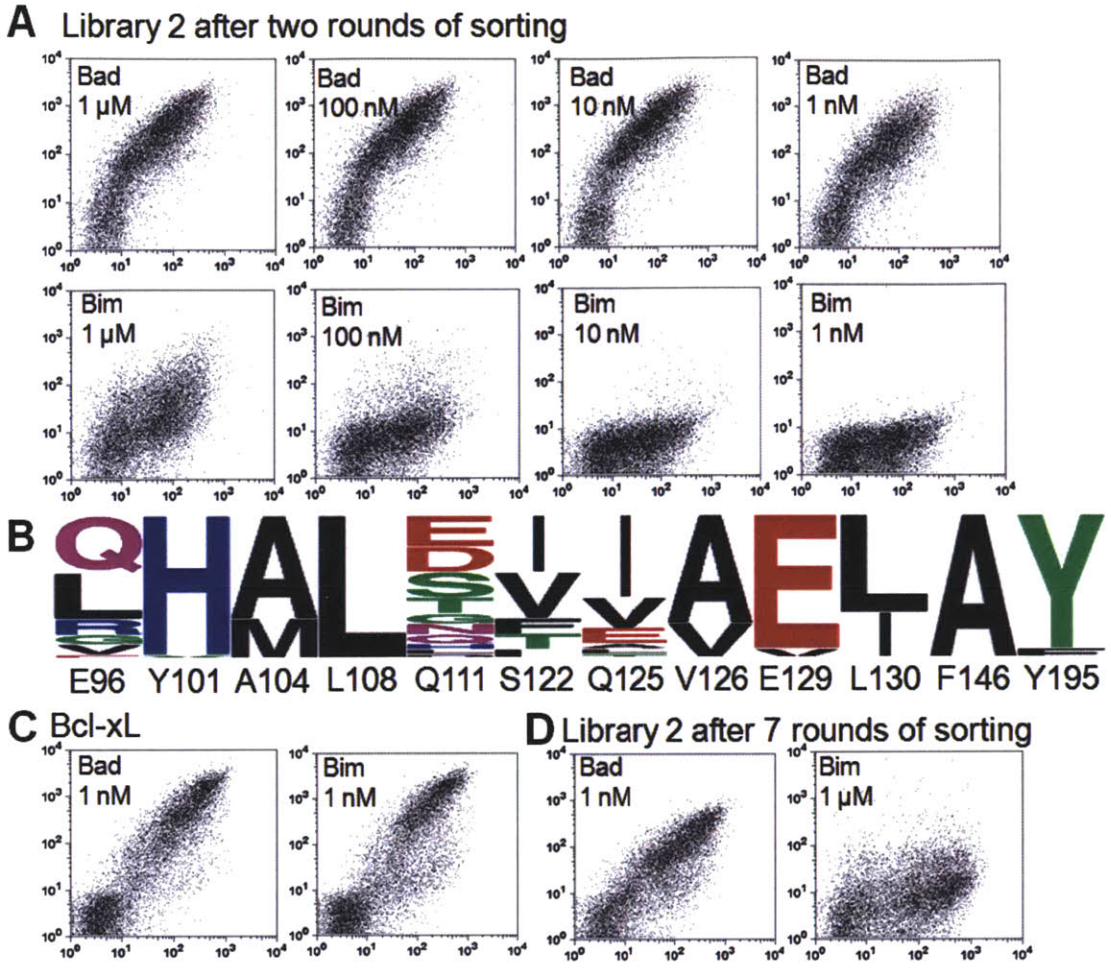


Figure 3.3 The second designed library
(A) Flow cytometry plots showing the population after two rounds of sorting of the second designed library displayed on the yeast surface binding Bad (top) and Bim (bottom) at 1 μM, 100 nM, 10 nM and 1 nM. (B) Sequence frequency plot for 28 unique sequences (Table 3.4) identified as specific for Bad over Bim BH3 from the second designed library. Mutations not located at the intended designed positions were omitted from the plot but are listed in Table 3.4. (C) Flow cytometry plots showing binding profiles of native Bcl-xL displayed on the surface of yeast toward Bad and Bim at 1 nM. Expression and binding signals are plotted on the x and y axes, respectively. (D) Flow cytometry plots showing binding of the second designed library after 7 rounds of sorting toward Bad at 1 nM and Bim at 1 μM.

Table 3.4 Unique sequences of specific binders isolated from the second designed library after two sorts [a]

	E96	Y101	A104	L108	Q111	S122	Q125	V126	E129	L130	F146	Y195	Other
A1	Q	H	M	L	D	V	V	V	E	I	A	Y	
A3	L	H	A	L	G	I	I	A	E	L	A	Y	
A4	Q	H	E	L	E	V	I	A	V	L	A	F	
A5	L	H	A	L	S	L	I	A	E	L	A	Y	
B1	L	H	M	L	E	F	I	A	E	L	A	Y	
B3	Q	H	A	L	T	I	L	A	E	L	A	Y	
B4	Q	H	A	L	S	V	I	V	E	I	A	Y	
B6	R	H	A	L	S	V	T	A	E	L	A	Y	
C2	V	H	A	L	T	I	I	A	E	L	A	Y	
C6	V	H	A	L	Q	I	E	A	E	L	A	Y	
D2	Q	H	A	L	D	V	V	V	E	I	A	Y	
D4	Q	H	A	L	D	I	I	V	E	I	A	Y	A85T
D5	G	H	A	L	S	V	V	A	E	L	A	Y	
E2	L	H	A	L	P	I	I	V	E	I	A	Y	F105I
E3	G	H	A	L	E	V	I	A	V	L	A	F	
E4	Q	Y	M	L	E	I	I	A	E	I	A	Y	
E6	Q	H	A	L	L	F	I	A	E	I	A	Y	
F1	L	H	A	L	Q	I	V	V	E	I	A	Y	
F2	E	H	M	L	E	V	I	V	E	I	A	Y	
F3	R	H	A	L	T	T	I	A	E	L	A	Y	
F4	R	H	A	L	H	V	V	A	E	L	A	Y	
F5	L	H	A	L	S	I	E	A	E	L	A	Y	
G2	L	H	A	L	D	T	E	A	E	L	A	Y	F105L
G3	L	H	M	L	E	F	I	A	E	L	A	Y	
G5	Q	H	M	L	D	T	A	A	E	L	A	Y	F105L
G6	L	H	A	L	G	L	I	A	E	L	A	Y	
H3	Q	H	M	L	N	I	V	A	E	L	A	Y	
H5	Q	H	M	L	N	I	I	A	E	L	A	Y	

[a] See descriptions for Table S1

Table 3.5 Sequences of clones from the final sorted population of the 2nd designed library

	E96	Y101[a]	F105[a]	Q111	S122I	Q125	V126	L130	A142	F146
A1	E	H	L	G	I	V	A	I	G	A
F1	L	H	L	D	T	E	A	L	G	A

[a] Y101H was not included in the modeling but was included in library 1 due to codon choice. Position 105 was not included in the designed library. All other residues were predicted to be non-disruptive.

Solution binding study

To confirm that the specificity profile of the selected Bcl-xL variants seen on the yeast surface could be recapitulated with soluble, recombinant proteins, we used a fluorescence polarization (FP) competition binding assay (see Materials and Methods, Fig 3.4, Table 3.7). Native Bcl-xL interacted very strongly with both the Bim and Bad 28-mer recombinant peptides (Bim-28 and Bad-28, Table 3.6), with fitted K_d values below 0.1 nM (Fig. 3.5A). The fitted K_d values for L2-7-A1 interacting with Bim-28 and with Bad-28 were 2.3 μM and 0.25 nM, respectively (Fig. 3.5B). The tightest binding that can be reliably quantified using our experimental conditions is ~0.1 nM. Thus, to more reliably measure the increase in specificity from native Bcl-xL to L2-7-A1 in solution, we turned to BH3 peptides of shorter length (22-mer) expected to be of lower binding affinity. The shorter peptides (Bim-22 and Bad-22, Table 3.6) maintained interactions with all of the designed Bcl-xL positions, based on crystal structures. The fitted K_d values of L2-7-A1 were > 50 μM and 33 nM, for Bim and Bad respectively, suggesting a specificity increase of >1,000 fold for the designed protein (Fig. 3.5C, Fig. 3.6).

Dissection of residues important for specificity

To analyze how individual mutations at each designed position contribute to the binding specificity of L2-7-A1, we made point mutations in Bcl-xL (Fig. 3.4, Fig. 3.5C) and also individually reverted selected residues of L2-7-A1 back to the native Bcl-xL amino acid (Table 3.8). We examined binding of these variants to Bim and Bad BH3 peptides. In the context of Bcl-xL (Fig. 3.5C), V126A, S122I and A142G preferred binding Bad over Bim, while L130I and F146A weakened binding to both peptides and showed no strong preference. Y101H, L105F and Q111G actually displayed preference for binding Bim over Bad. When examined in the context of L2-7-A1 (Table 3.8), reverting I122 back to Ser, G142 back to Ala, and surprisingly, A146

back to Phe all caused significant loss of Bad over Bim specificity. The loss in specificity for L2-7-A1-A146F was particular interesting as it likely explained why the F146A mutation was present in all specific sequences in library 2 (Fig. 3.3B) but did not confer specificity when measured in the context of Bcl-xL (Fig. 3.5C). Interestingly, reverting the two mutations not intended to be included as non-disruptive residues in the library, Y101H and L105F, caused moderate loss in specificity as well (L2-7-A1-H101Y and L2-7-A1-L105F in Table 3.8), despite favoring Bim binding over Bad when made in the context of Bcl-xL (Fig. 3.5C). Overall, the analysis suggested that although some of the influences of the designed residues are relatively independent of the sequence context, significant higher-order interaction among residues at different positions was evident and contributed to the observed specificity.

Table 3.6 BH3 peptides used in this study

	efgabcdefgabcdefgabcdefgab	Origin[a]
Bim-28	MRPEIWIAQELRRIGDEFNAYYARRVFL	recombinant
Bad-28	LWAAQRYGRELRRMSDEFVDSFKKGLPR	recombinant
Bim-22	MRPEIWIAQELRRIGDEFNAYY	synthetic
Bad-22	LWAAQRYGRELRRMSDEFVDSF	synthetic
Bak	SSTMGQVGRQLAIIGDDINRRYDSEFQT	recombinant
Bax	DASTKKLSECLKRIGDELDSNMELQRMI	recombinant
Beclin	GGTMENLSRRLKVTGDLFDIMSGQTDVD	recombinant
Bid	EDIIRNIARHLAQVGDSMDRSIPPGLVN	recombinant
Bik	MEGSDALALRLACIGDEMDVSLRAPRLA	recombinant
Bmf	HQAEVQIARKLQCIADQFHRLHVQQHQQ	recombinant
Hrk	SSAAQLTAARLKALGDELHQRTMWRRRA	recombinant
Mule	GVMTQEVGQLLQDMGDDVYQQYRSLTRQ	recombinant
Noxa	AELEVECATQLRRFGDKLNFRQKLLNLI	recombinant
Puma	EQWAREIGAQLRRMADDLNAQYERRRQE	recombinant
fBad-21[b]	NLWAAQRYGRELRRMSDKFVD	synthetic
fBad-23[c]	Fl-NLWAAQRYGRELRRMSDEFVDSF	synthetic
fBad-27[c]	Fl-NLWAAQRYGRELRRMSDEFVDSFKKGL	synthetic

[a]A detailed description of the constructs is included in Materials and Methods.
[b]The FITC label is on the underlined Lys residue
[c] Fl stands for FITC

Table 3.7 Fitted K_d values for direct binding experiments between Bcl-xL variants and different fluorescently labeled peptides (fitted curves shown in Fig. 3.4)

	Labeled peptide	K_d (nM)
Bcl-xL	fBad-23	3.3
L2-7-A1	fBad-23	22
Y101H	fBad-23	23
A104M	fBad-21	3.8
F105L	fBad-23	8.7
Q111G	fBad-23	6.5
S122I	fBad-21	4.2
Q125V	fBad-21	4.6
V126A	fBad-23	15
L130I	fBad-23	31
A142G	fBad-21	1.5
F146A	fBad-23	61
L2-7-A1-H101Y	fBad-23	2.9
L2-7-A1-L105F	fBad-27	6.8
L2-7-A1-G111Q	fBad-23	13
L2-7-A1-I122S	fBad-27	40
L2-7-A1-G142A	fBad-27	34
L2-7-A1-A146F	fBad-23	21

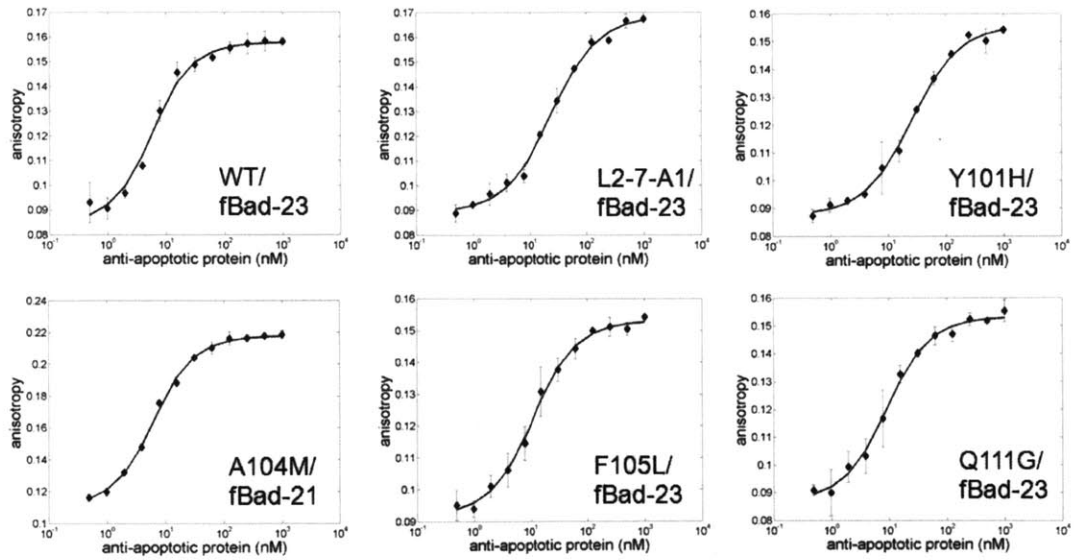


Figure 3.4 Fluorescence polarization experiments and fitted curves characterizing binding of Bcl-xL, L2-7-A1 and different point mutants to fluorescently labeled Bad peptides. Different labeled peptides used for different Bcl-xL variants were indicated on the plot and described in Table 3.6. The averaged values as well as the spread of two independent measurements of anisotropy were plotted as a function of Bcl-xL variant concentration. Experimental conditions and curve fitting for the direct binding experiments are described in Materials and Methods. Figure to be continued on the next page.

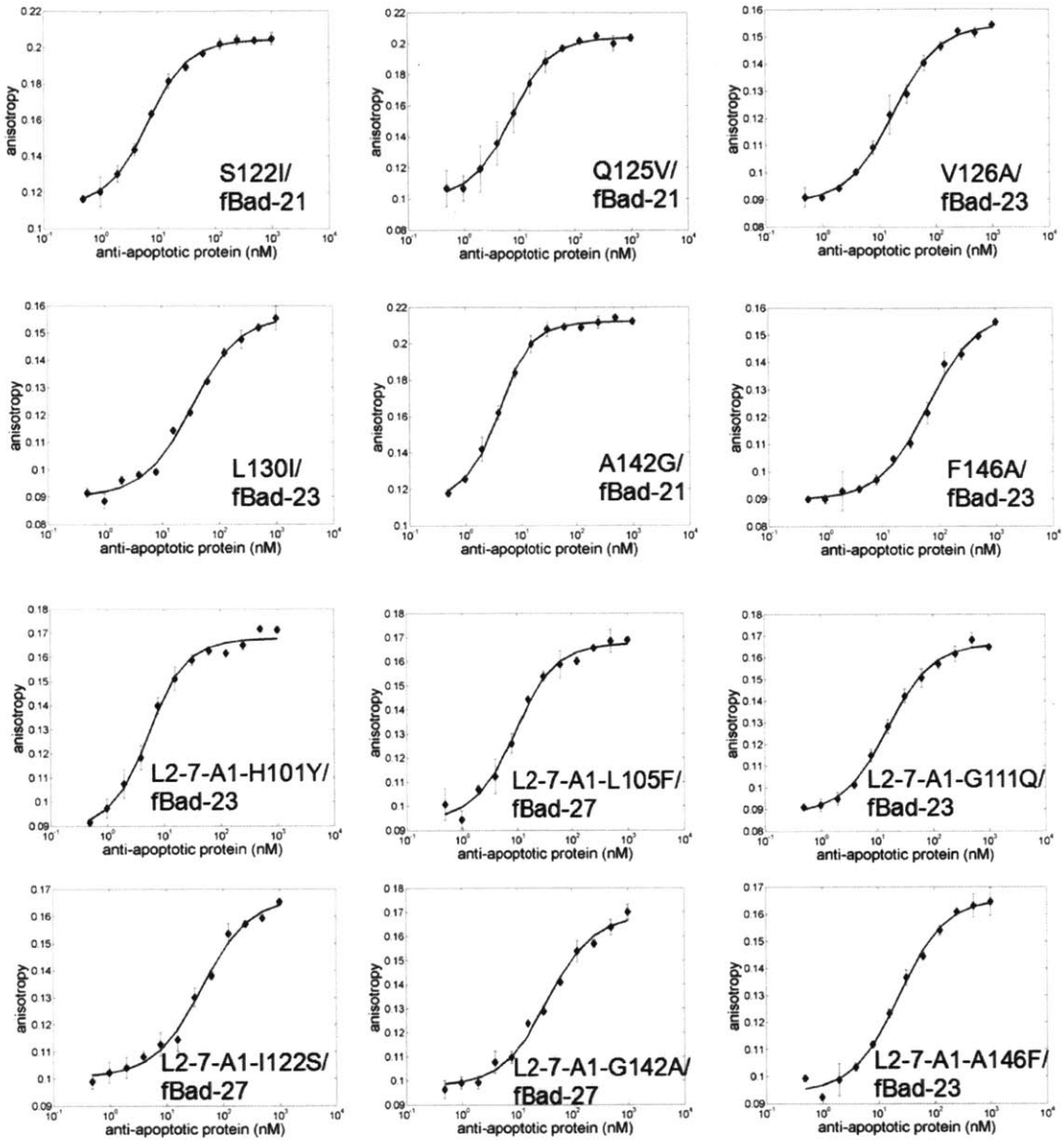
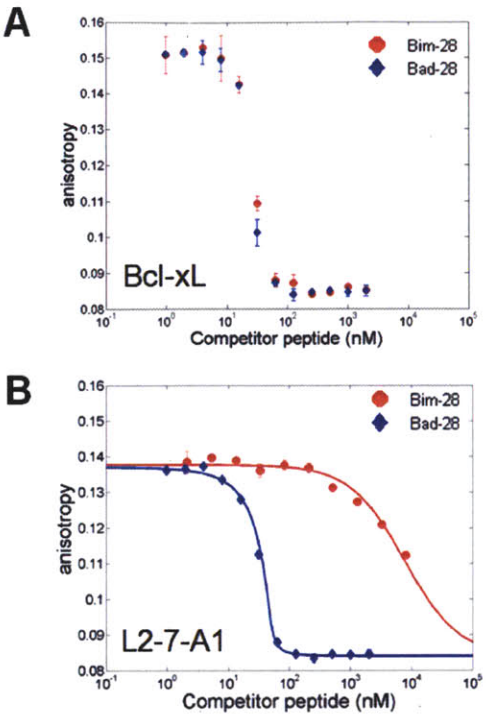


Fig 3.4 (continued) Fluorescence polarization experiments and fitted curves characterizing binding of Bcl-xL, L2-7-A1 and different point mutants to fluorescently labeled Bad peptides.



C

	Bim-22	Bad-22
Bcl-xL	3.2	1.8
L2-7-A1	> 50,000	33
Y101H	10	25
A104M	**2.3**	**0.44**
F105L	4.1	8.4
Q111G	2.0	10.5
S122I	**1.3**	**0.17**
Q125V	0.66	0.39
V126A	**98**	**20**
L130I	26	38
A142G	**2.7**	**0.15**
F146A	75	67

Figure 3.5 Fluorescence polarization experiments characterizing Bcl-xL and its variants binding to BH3 peptides derived from Bim or Bad.
(A) Competition of Bim-28 and Bad-28 with fBad-23 in binding to native Bcl-xL. Binding conditions are described in Materials and Methods, and peptide sequences are given in Table S1. Fitted curves were not shown as the interactions were too tight to be fitted. (B) Competition of Bim-28 and Bad-28 with fBad-23 in binding to the design L2-7-A1. For (A) and (B), the averaged values as well as the spread of two independent measurements were plotted as a function of competitor peptide concentration. (C) K_d values for Bcl-xL, L2-7-A1 and different Bcl-xL point mutants interacting with Bim-22 and Bad-22. Mutations that give the most Bad-over-Bim specificity are highlighted in blue (The mutation A104M was not present in L2-7-A1). The curves are shown in Fig. 3.5, and the fluorescent peptides being competed off are indicated in Table 3.7.

Table 3.8 K_d values for point mutants of design L2-7-A1 binding Bim/Bad

	Bad-22	Bad-28	Bim-28
L2-7-A1	33	0.25	2,300[a]
L2-7-A1-H101Y	4	-	200[a]
L2-7-A1-L105F	-	1.3	1,500[a]
L2-7-A1-G111Q	17	-	1,400[a]
L2-7-A1-I122S	-	7.0	3,200[a]
L2-7-A1-G142A	-	7.6	430
L2-7-A1-A146F	39	-	130

[a]Limited solubility of Bim-28 at high concentration prevents accurate determination of the lower baseline when fitting dissociation constants. Fitted values were obtained instead by imposing a lower baseline as described in Materials and Methods.

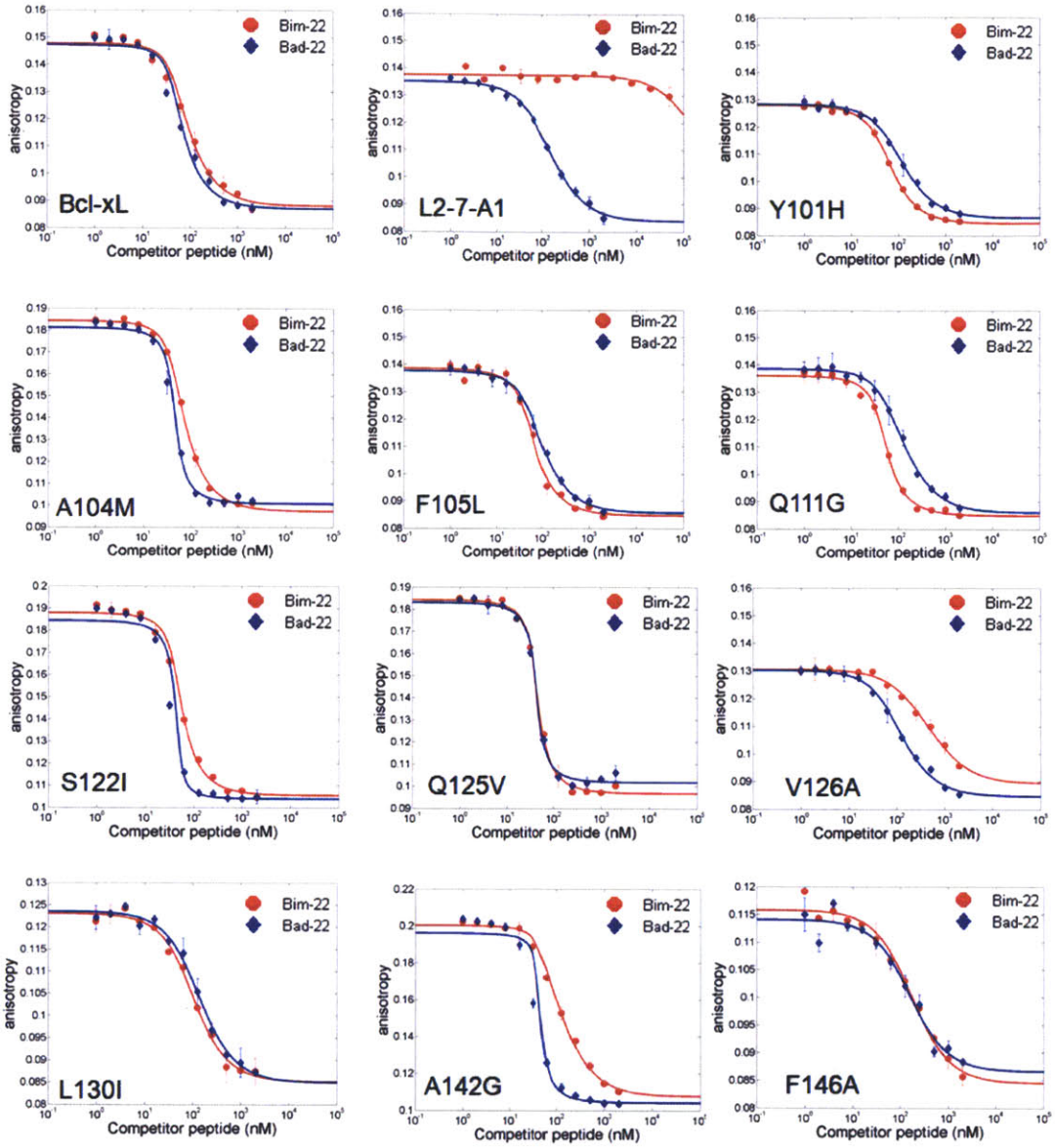


Figure 3.6 Fluorescence polarization experiments and fit curves characterizing binding of Bcl-xL, L2-7-A1 and different point mutants to unlabeled Bim or Bad peptides by competition.

Different fluorescently labeled peptides being competed off of different Bcl-xL variants were shown in Table 3-6. The Bcl-xL or L2-7-A1 variant and the competitor peptides were indicated on the plot. The averaged values as well as the spread of two independent measurements were plotted as a function of competitor peptide concentration. Experimental conditions and curve fitting for these competition experiments are described in Materials and Methods. Figure to be continued on the next page.

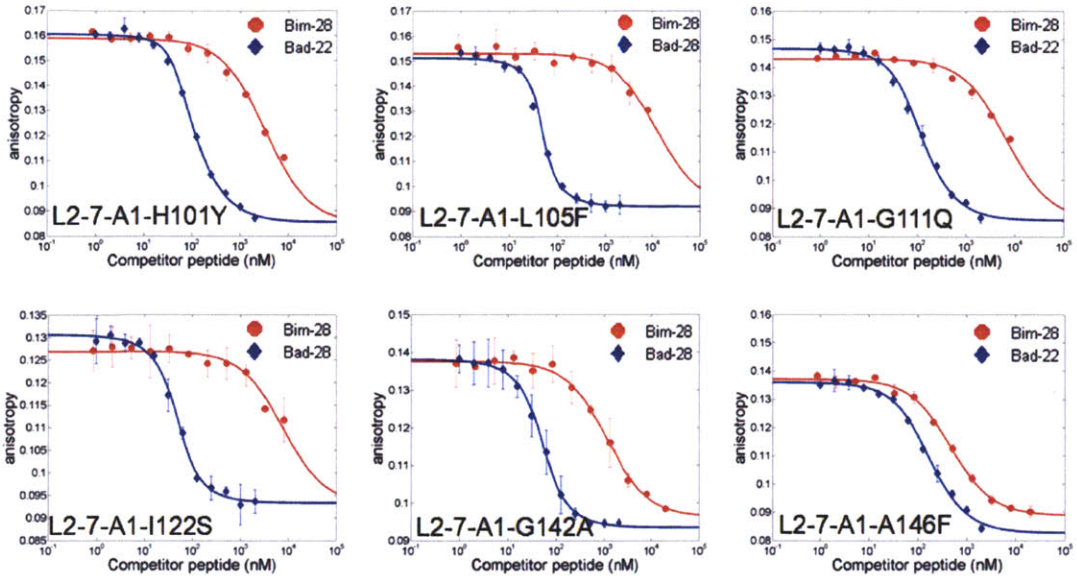
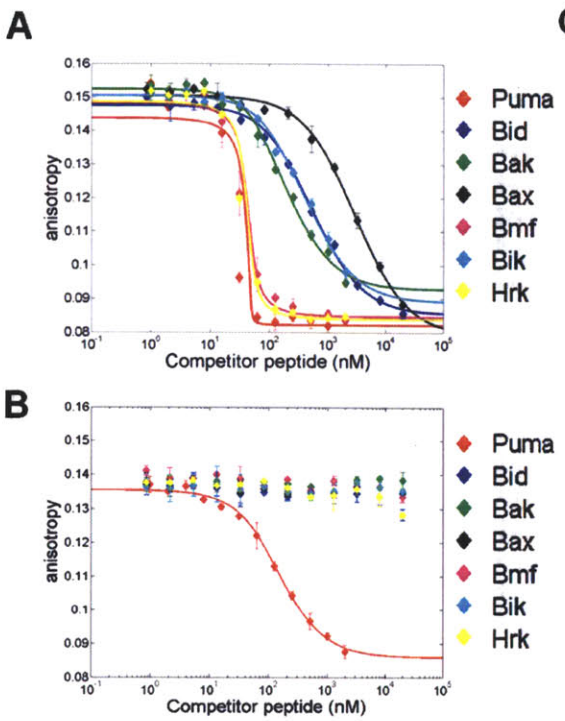


Figure 3.6 (continued) Fluorescence polarization experiments and fit curves characterizing binding of Bcl-xL, L2-7-A1 and different point mutants to unlabeled Bim or Bad peptides by competition.

Specificity profiles against other BH3s

We also evaluated interactions between L2-7-A1 and 10 other peptides derived from the BH3 regions of human Bcl-2 family proteins not included in the design/screening efforts. In contrast to Bcl-xL, which interacts strongly with several other BH3s (Fig. 3.7A, 3.7C), significant interaction was observed only between L2-7-A1 and PUMA (Fig. 3.7B). The interaction of L2-7-A1 with PUMA was significantly weaker than that between Bcl-xL and PUMA (Fig. 3.7C). In summary, L2-7-A1 displayed global specificity against the other BH3s not included in specificity screening.



C

	Bcl-xL	L2-7-A1
Bad	<0.1	0.25
Bim	<0.1	2,300
PUMA	<0.1	35
Bid	33	>20,000
Bak	11	>20,000
Bax	200	>20,000
Bmf	0.24	>20,000
Bik	25	>20,000
Hrk	0.17	>20,000
Beclin	>1,000	>20,000
MULE	>20,000	>20,000
Noxa	>20,000	>20,000

Figure 3.7 Fluorescence polarization experiments characterizing Bcl-xL and the design L2-7-A1 binding to 10 native BH3 peptides.
(A) Competition of different BH3s with fBad-23 binding to native Bcl-xL. (B) Competition of different BH3 peptides with fBad-23 binding to the design L2-7-A1. For (A) and (B), the averaged values as well as the spread of two independent measurements were plotted as a function of competitor peptide concentration. Competition curves were shown only for competitor peptides binding Bcl-xL significantly. (C) Fitted K_d values of Bcl-xL and L2-7-A1 interacting with different BH3 peptides.

Discussion

The idea underlying our library design strategy was that maintaining high useful sequence diversity is important when treating difficult design problems such as protein-protein interaction specificity. The importance of diversity has been illustrated and discussed previously in other contexts[17,22]. In our approach, diversity is obtained by including predicted non-disruptive residues in addition to specific residues at each designed position. We enforced the inclusion of all predicted specific residues in our libraries, thereby providing access to a variety of possible predicted specificity strategies. We also adjusted the inclusion of non-disruptive residues to achieve the desired library size.

We compared the library screening outcomes and the behaviors of the Bcl-xL point mutants with structural modeling results. Our library design procedure was successful in enriching the library in Bad binders; compared to our prediction that 6% of library 1 lacked any disruptive mutations, 5% of expressed library 1 sequences were observed to bind to 10 nM Bad (Fig. 3.2A). Nonetheless, the performance of the specificity predictions suggested room for improvement in the modeling protocol, which was tailored to be very inclusive in picking non-disruptive residues. All 9 mutations in the selected L2-7-A1 sequences were characterized as Bcl-xL point mutants (Fig. 3.5C). The mutation S122I was predicted and confirmed to be specific. Two other specific mutations, V126A and A142G, were included in the library only as predicted non-disruptive residues. G111Q and L130I were predicted but shown not to be specific. Two other non-specific mutations, Q125V and F146A, were also predicted only as non-disruptive residues. However, F146A likely played an important role in specificity as reversing the mutation in L2-7-A1 caused significant loss in specificity (Table 3.8). Y101H was not selected for modeling and F105L was not included in the library. Neither of them was specific when characterized as Bcl-xL point

mutants, but might contribute to specificity when combined with other residues, as observed when examined in the context of L2-7-A1. For predicted specificity residues not present in L2-7-A1, A104M was confirmed to be specific and appeared in some of the specific sequences. However, most of the other predicted specificity residues were not present in the selected specific sequences.

Given that many of the predicted specificity resides were in fact not specific, the broad inclusion of non-disruptive residues in this application was crucial for capturing important residues missed by the specificity predictions (e.g. V126A, A142G and F146A). The strategy we chose had the caveat that many residues predicted to be non-disruptive would indeed not contribute to specificity. In fact, as shown in Fig. 3.2A, the designed library showed on average stronger binding to Bim than to Bad. Adjusting the inclusion of non-disruptive residues by the desired library size was therefore important. We picked a conservative library size of 10^7 for yeast surface display. However, the ILP optimization makes it possible to examine how relaxing the library size constraint, with the risk of experimentally under-sampling the diversity, could lead to a more comprehensive inclusion of the non-disruptive residues. One of the advantages of the ILP formulation is that it allows the designer to explore trade-offs up front in this way, knowing that the algorithm provides optimal solutions under the specified constraints.

Mutational analysis was also used to understand the origins of the high specificity of L2-7-A1 for Bad over Bim (Fig. 3.5C). Several single mutations conferred moderate to strong Bad-over-Bim specificity (S122I, V126A and A142G), whereas others significantly weakened Bim binding (Y101H, L130I and F146A). However, this explains only part of the specificity, and analysis of selected L2-7-A1 mutants (Table 3.8) suggested that inclusion of higher order interactions is likely needed to understand the complete picture as described in Results. If

reliable predictions of higher-order coupling could be made, including them in library design could increase efficiency. For example, Lippow et al. suggested the importance of explicitly looking at higher-order interactions in the context of enzyme design[20]. Coupling between codons at different positions can be easily incorporated under the ILP formulation if desired[22].

Interestingly, design L2-7-A1 is not only specific against Bim, but also against all other natural BH3s tested in this study. The only other BH3 peptide that showed significant interaction with L2-7-A1 was PUMA, which shares some features with Bad, such as Met at position 3d and Ala (closer in size to Ser in Bad) at 3e (Table 3.6). Design examples where specificity was obtained "for free", i.e. without explicit consideration, have been reported previously[34,35]. In the present case, specificity against Bim was not "free" but had to be introduced by screening; the original library 1 bound strongly to both Bad and Bim. Elements that destabilize interaction with Bim apparently also destabilize interaction with many other BH3 peptides. For challenging multi-specificity design problems where it is impractical to screen against all relevant competitors it might be reasonable to take the approach used here, as long as designed positions are selected carefully to reach meaningful diversity as described before. An interesting analogy is the study performed by Guntas et al.[36], which showed that a library enriched in predicted well-folded sequences performed as well as one enriched in predicted binders when screening for novel interaction partners. In contrast, design studies targeting bZIP coiled coils showed that ignoring some competitors in design calculations could lead to undesired binding[37]. The degree to which negative design is required appears to depend critically on the particular problem being addressed[38].

Materials and Methods

Structural modeling

Structural models of Bcl-xL point mutants interacting with Bim or Bad were generated using Rosetta 3.0[12]. The crystal structure of human Bcl-xL in complex with Bim (PDB ID: 3FDL)[32] was used to model interactions between Bcl-xL mutants with Bim and Bad, and that of mouse Bcl-xL in complex with Bad (PDB ID: 2BZW)[33] was used to model interactions between Bcl-xL mutants with Bad only. An ensemble of 100 structures was derived separately from each of 3FDL and 2BZW, with fixed native sequence, using the backrub flexible-backbone modeling utility in Rosetta[39]. The entire binding interface was allowed to change structure during the backrub sampling. Each Bcl-xL mutant interacting with Bim or Bad was then modeled on all members of the structural ensemble using the fixed backbone design mode in Rosetta. A 50-step conjugate-gradient based minimization was performed for each ensemble member, and the Rosetta energy for each minimized structure within the ensemble was obtained. The minimum energy was defined as the score of the interaction between the Bcl-xL mutant being modeled and Bim or Bad, and the difference relative to the score of native Bcl-xL interacting with Bim or Bad was calculated (ΔE_{Bim} or ΔE_{Bad}). The unbound states were not modeled and the 20 single amino acid reference energies in Rosetta were used as the reference state instead. The score should therefore not be viewed as an attempt to predict whether the mutant would increase or decrease the affinity of the interaction, but rather as a simple metric estimating if complex formation would be greatly disrupted. Note that it is possible that a mutant could be predicted to disrupt significantly the energy for both the complex and the unbound state. Such mutant would be predicted to not disrupt the interaction and be missed by our protocol. On the other hand, mutants predicted to strongly destabilize the unbound states might not be desired. As interactions

between mutants with Bad were modeled using both 3FDL and 2BZW as templates, two values of ΔE_{Bad} were generated and the lower one was picked as the final ΔE_{Bad}. Residues with ΔE_{Bad} lower than 3 (for the first designed library) or 1 Rosetta energy unit (for the second designed library) were defined as non-disruptive residues. Residues with ΔE_{Bim} - ΔE_{Bad} greater than 2 (for the first designed library) or 3 Rosetta energy units (for the second designed library) were defined as specificity residues. Here the omission of the unbound states should not influence the specificity prediction as they would be canceled out regardless.

Position Y195 was not subjected to structural modeling as it was missing in the human Bcl-Xl/Bim structure (3FDL). The corresponding position (Y195) was observed in the mouse Bcl-xL/Bim structure (1PQ1) and formed a hydrogen bond with N102 (position 4b in the BH3 alignment) of Bim (occupied by Val in Bad). The manual choice of Phe was included to explore whether removing this hydrogen bond would provide specificity.

Selecting degenerate codons for the designed library

At each designed position *i*, we defined two quantities for each degenerate codon *j*: (1) the size, s_{ij}, which is the number of unique trinucleotides within codon *j*. (2) n_{ij}, the number of "non-disruptive" residues encoded by codon *j*. The codons were pre-filtered by the following two criteria: (1) The native amino acids and all "specificity residues" at the position must be encoded by the codon. (2) Codons encoding only the native amino acid were eliminated. (3) Among the pool of degenerate codons passing the first two criteria, any codon with a larger s_{ij} but a smaller n_{ij} than another within the pool at position *i* was eliminated. This process was repeated for every pair of codons until no codon was "dominated" by another. Optimization of degenerate codon combinations, out of the remaining pool of codons J_i at each designed position *i*, was performed by solving the following integer linear programming problem:

$$\text{Max } \sum_i \sum_{j \epsilon Ji} c_{ij} \log(n_{ij}), \text{ under } \sum_i \sum_{j \epsilon Ji} c_{ij} \log(s_{ij}) <= 7, \text{ and } \sum_{j \epsilon Ji} c_{ij} = 1 \text{ for each position } i$$

Where $c_{ij} = 1$ if codon j was picked at position i, and 0 otherwise. For the winner codon j picked at each position i, $\sum_i \log(n_{ij}) = \log(\prod_i n_{ij})$ is the logarithm of the number of unique protein sequences encoded with all designed positions occupied by non-disruptive residues, and $\sum_i \log(s_{ij}) = \log(\prod_i s_{ij})$ is the library size (or the number of unique DNA sequences in the library) as described in the text. The problem was solved using the glpsol solver in the GLPK package (GNU MathProg). Note that occasionally multiple codons at one position could have identical statistics and all be optimal under this formulation, and in this case we manually examined the choices and selected one codon.

To design the 2nd library, a "degenerate codon pair" was considered in addition to individual degenerate codons at each designed position. This provided greater flexibility in sampling desired sequence features within a fixed library size. A "degenerate codon pair" was defined as two degenerate codons that are orthogonal to each other, i.e. there is no overlap between the tri-nucleotides specified by the two codons. Experimentally, a designed position can be constructed as a codon pair simply by mixing oligonucleotides, or by mixing the PCR products generated by using each of two individual degenerate codons at a site (chosen in this study). The size, s_{ij}, for a pair j is the sum of its two codon components, and the experimental mixing ratio is simply the ratio of the sizes for the two codons. The metric n_{ij} is the total number of unique non-disruptive residues from the two components. The filtering process as described above can be performed for a "codon pair" separately from normal codons. We imposed an additional filtering criterion to exclude any stop codons. For the optimization process, each designed position could be encoded as a single degenerate codon or a pair of codons. However, to avoid an explosion of steps in the library construction protocol, additional constraints were imposed to ensure that no

more than one designed position was encoded by a "codon pair" within the same oligonucleotide in the PCR based assembly procedure (Table 3.9)

$$\sum_i \sum_{j \epsilon Ji} c_{ij} p_{ij} <= 1 \text{ for all position } i \text{ randomized on the same oligonucleotide}$$

where p_{ij} is 1 if j is a codon pair at position i and 0 otherwise. The criterion for deciding which designed positions were encoded by the same oligo was described later. The same ILP optimization procedure was then solved with the above constraints to obtain the second designed library.

Cloning, protein expression and purification

For yeast surface display, the human Bcl-xL gene (1-209), followed by a GGGGSG linker and a C-terminal myc tag (give sequence), was cloned into the pCTCON2 vector via NheI and BglII sites, with the gene fused in frame to the C-terminus of Aga2p with a $(GGGGS)_3$ linker. PCR amplification of the Bcl-xL gene was performed using a previously made MBP Bcl-xL fusion as the template. For recombinant proteins used in the fluorescence polarization assay, the Bcl-xL gene and variants obtained from screening were cloned into a modified pDEST17 vector via BamHI and XhoI sites. A BamHI cut site was present in the Bcl-xL gene, and therefore either a BglII or a BclI site, both compatible for ligation to a BamHI cut vector, was included in the primers for PCR amplification. Mutants of either the Bcl-xL gene or the L2-7-A1 design were made using PCR based sited directed mutagenesis followed by blunt end ligation[40], or Quick change (Agilent). Recombinant human BH3 peptides (Bim-28, residues 142-169; Bad-28, residues 104-131; other BH3 28-mers, Table 3.6), with a C-terminal GG linker followed by a Flag tag sequence (DYKDDDDK), were constructed by gene synthesis. Primers were designed using DNAWorks[41], and a two-step PCR procedure was used for annealing and amplification.

The genes were then cloned into a modified pDEST17 vector via BamHI and XhoI sites. Recombinant Bcl-xL proteins and BH3 peptides (with a His_6 tag) were expressed in *E. coli* RP3098 cells. Culture was grown at 37 ℃ until OD ~0.4-0.9, and expression was induced by addition of 1 mM IPTG. Purification of Bcl-xL proteins was performed under native condition using Ni-NTA. An additional step of gel-filtration purification with a HiLoad Superdex^{TM} 75 column (GE) was performed for the mutants and the designed proteins because protein oligomerization was observed for some of them. Purification of BH3 peptides was performed under denaturing condition using Ni-NTA and followed by reverse-phase HPLC and their masses subsequently verified by MALDI spectrometry.

Making combinatorial libraries

The oligonucleotides introducing diversity for the two designed libraries are shown in Table 3.9. PAGE-purified oligonucleotides were ordered from Integrated DNA Technology. Two randomized positions were chosen to be encoded by the same oligo if the length of the constant region between them is shorter than 15 nucleotides. The protocol to introduce degenerate codon pairs (applicable only to the second designed library) is described under the library design section. The first library was constructed by PCR overlap extension joining two PCR fragments, #1-1 and #1-2. Fragment #1-2 was PCR amplified from the PCR fragment #1-2a. PCR amplification for fragment #1-1 introduced diversity for positions 97, 101, 104, 108 and 112 using Bcl-xL in pCTCON2 as the template. PCR amplification for fragment #1-2a randomized positions 126, 129 and 130 using Bcl-xL in pCTCON2 as the template, and subsequent PCR amplification for fragment #1-2 randomized position 142 using fragment #2a as the template.

The second designed library was made similarly using PCR overlap extension joining two PCR fragments, #2-1 and #2-2. Fragment #2-1 was PCR amplified from the PCR fragment #2-1a,

and fragment #2-2 was made using PCR overlap extension joining PCR fragment #2-2a and PCR fragment #2-2b. Fragment #2-2b was PCR amplified from the PCR fragment #2-c. PCR amplification for fragment #2-1a introduced diversity for positions 96, 101, 104, 108 and 111 using Bcl-xL in pCTCON2 as the template. PCR amplification for fragment #2-1 introduced diversity for positions 122, 125, 126, 129 and 130 using fragment #2-1 as the template. PCR amplification for fragment #2-2a introduced diversity for positions 146 and 195 using clone C1 from the first designed library (in the pCTCON2 vector) as the template.

The final PCR products were co-transformed with pCTCON2 vector, cut with NheI/XhoI, into yeast following the procedure of Chao et al.[42] using a BioRad Gene Pulser.

Yeast surface display, flow cytometry analysis and cell sorting

Yeast strain EBY100 and the plasmid for yeast surface display (pCTCONT2) were a generous gift from Dr. K. D. Wittrup (Massachusetts Institute of Technology). Yeast cells were grown overnight at 30 °C in SDCAA media, and display of the Bcl-xL protein was induced by switching to SGCAA media for > 12 hr following protocols described by Chao et al.[42]. Induced cells were washed with TBS (50 mM Tris, 100 mM NaCl, pH 8.0), and incubated with different concentrations of Bim-28 or Bad-28 for 1-2 hr in TBS at ~25 °C. Cells were then washed with cold TBS and labeled with primary antibodies (anti-c-*myc* rabbit and anti-His mouse, Sigma) at 1:67 (anti-c-*myc*) or 1:100 (anti-His) dilution for 30 min - 2 hr in BSS (TBS with 1 mg/mL bovine serum albumin) at 4 °C. After washing in cold BSS, cells were labeled with secondary antibodies (PE conjugated anti-rabbit, Sigma and APC conjugated anti-mouse, BD Bioscience) at 1:100 dilution for 30 min – 2 hr in BBS at 4 °C. Cells were then washed again in cold BBS prior to analysis or sorting. The analysis was performed on BD FACSCalibur-HTS1 (BD Bioscience), and the sorting on BD FACSAria (BD Bioscience) or MoFlo (Beckman Coulter).

Cells were gated by forward light scattering to avoid the analysis/sorting of clumped cells. Data were analyzed using FlowJo (Tree Star, Inc.).

Below we described sorting of the first and the second designed libraries in more detail. The first designed library was subjected to one round of positive sorting (gating for expression and binding) for cells binding 10 nM Bad-28, one round of positive sorting for cells binding 10 nM Bad-28 in the presence of 1 μM unlabeled Bim, two rounds of negative sorting (gating for expression without binding) against cells binding 10 nM Bim-28, one round of negative sorting against cells binding 100 nM Bim-28, and finally one round of positive sorting for cells binding 10 nM Bad. The second designed library was subjected to two rounds of positive sorting for cells binding 1 nM Bad-28 in the presence of 1 μM unlabeled Bim, two rounds of positive sorting for cells binding 1 nM Bad-28 in the presence of 5 μM unlabeled Bim, one round of negative sorting against cells binding 3 μM Bim-28, one round of negative sorting against cells binding 5 μM Bim-28, and finally one round of positive sorting for cells binding 1 nM Bad-28.

Generation of sequence frequency plot

For the first designed library, 48 individual clones from the final sorted population were examined for binding to 1 nM Bad-28 or 10 nM Bim-28. Twenty-one clones with unique sequences showed stronger binding signal for 1 nM Bad-28 over 10 nM Bim-28, and the sequence frequency plot shown in Fig. 3.2C was generated from these sequences (Table 3.2). For the second designed library, 48 individual clones from the population after two round of sorting were examined for binding to1 nM Bad-28 or 500 nM Bim-28. Twenty-eight clones with unique sequences (Table 3.4) showed stronger binding signal for 1 nM Bad-28 over 500 nM Bim-28 and were used to generate the frequency plot in Fig. 3.3B.

Fluorescence polarization binding assays

Unlabeled and FITC-labeled peptides (Table 3.6) were synthesized by the MIT Biopolymers Facility at the Koch Institute for Integrative Cancer Research. A purified 21-mer Bad peptide with a FITC-labeled lysine[43] was ordered from Calbiochem (now EMD Biosciences). Labeled peptides were ordered with free C-termini, and unlabeled peptides were ordered with free N and C-termini for enhanced solubility. Synthesized peptides were purified by reverse phase HPLC using a C18 column. All assays were performed in assay buffer (50 mM NaCl, 20 mM Na_2HPO_4, 1 mM EDTA, 0.01% Triton X-100, and 5% DMSO)[43] at ~25 °C. For direct binding assays, the concentration of the fluorescently labeled peptide was fixed at 5 nM. Serial dilution of the Bcl-xL protein or its variants was performed before mixing with the fluorescently labeled peptide. The reaction was allowed to equilibrate for at least 1 hr. For competition assays, the concentration of the fluorescently labeled peptide was fixed at 15 nM, and the Bcl-xL protein or its variant was fixed at 50 nM. Serial dilution of the unlabeled peptide was performed, before adding the mixture of fluorescently labeled peptide and the Bcl-xL protein or its variant. The reaction was allowed to equilibrate for at least 3 hr. Different fluorescently labeled peptides were used for experiments involving different Bcl-xL protein variants in order to obtain K_d values that could be fitted reasonably (Fig 3.4, Table 3.7). Non-binding 96 well plates (Corning Incorporated) were used for all assays. Anisotropy measurements were performed on a SpectraMax M5 (Molecular Devices) plate reader. All measurements were done in duplicates. The averaged values were plotted, along with error bars signifying the spread of the two measurements. Complete models for fitting K_d values for both direct binding and competition experiments were described before[44], and the K_d values were fit using Matlab (Mathworks). A lower baseline corresponding to the measured anisotropy value of the free fluorescently labeled

peptide in solution was enforced in the fitting for competition experiments for competitor peptides failing to reach close to complete inhibition at the highest concentration.

Table 3.9 Oligonucleotides introducing randomization

	Oligonucleotides	Positions randomized
1st library	5'-GCTGTCCCTGGGGTGATGTGCAHCTGGGATGTAVBGTCA CTGAAMNHCCGCCGAWRCCGCAGTTCMAWCTCGTCGCCTG CCTCCCTC-3' (reverse)	F97, Y101, A104, L108, L112
	5'-CACATCACCCCAGGGACAGCATATCAGAGCTTTGAACAG RBKGTGAATRHAVTCTTCCGGGATGGGGTAAACTGG-3'	V126, E129, L130
	5'-TTCCGGGATGGGGTAAACTGGGGTCGCATTGTGRSCTTTT TCTCCTTCGGCGGGGCAC-3'	A142
2nd library	5'-CTGTCCCTGGGGTGATGTGCAAMNBGGATGTCASGTCACT GAATGCCCGCCGATRCCGCAGTTCAAATHSGTCGCCTGCCTC CCTCAGC-3' (reverse)	E96, Y101, A104, L108, Q111
	5'-CTGTCCCTGGGGTGATGTGCAAMNBGGATGTCASGTCACT GAACATCCGCCGATRCCGCAGTTCAAATHSGTCGCCTGCCTC CCTCAGC-3' (reverse)	
	5'-CGACCCCAGTTTACACCGTCCCGGAAGAKTWCATTCACT RCADHTTCAAAGDNCTGATATGCTGTCCCTGGGGTGATGTG CAA-3' (reverse)	S122, Q125, V126, E129, L130
	5'-CGACCCCAGTTTACACCGTCCCGGAAGAKTWCATTCACT RCTHSTTCAAAGDNCTGATATGCTGTCCCTGGGGTGATGTG CAA-3'	
	5'-CTTCCGGGACGGTGTAAACTGGGGTCGCATTGTGGGCTTT TTCTCCTTSGGCGGGGCACTGTGCGTGG-3'	F146
	5'-CTTCCGGGACGGTGTAAACTGGGGTCGCATTGTGGGCTTT TTCTCCGCTGGCGGGGCACTGTGCGTGG-3'	F146
	5'-CTCTCGGCTGCTGCATTGTTCCCGWAGAGTTCCACAAAAG TATCCCAGC-3' (reverse)	Y195

Acknowledgements

We thank the staff at the M.I.T flow cytometry facility for assistance, and R. Cook and the M.I.T Biopolymers facility for peptide synthesis. We thank members of the Keating lab for helpful discussions and the Bell lab for use of equipment. This work was funded by NIGMS awards GM084181.

References

1. Steed, P. M., Tansey, M. G., Zalevsky, J., Zhukovsky, E. A., Desjarlais, J. R., Szymkowski, D. E., Abbott, C., Carmichael, D., Chan, C., Cherry, L., Cheung, P., Chirino, A. J., Chung, H. H., Doberstein, S. K., Eivazi, A., Filikov, A. V., Gao, S. X., Hubert, R. S., Hwang, M., Hyun, L., Kashi, S., Kim, A., Kim, E., Kung, J., Martinez, S. P., Muchhal, U. S., Nguyen, D. H., O'Brien, C., O'Keefe, D., Singer, K., Vafa, O., Vielmetter, J., Yoder, S. C. & Dahiyat, B. I. (2003). Inactivation of TNF signaling by rationally designed dominant-negative TNF variants. *Science* **301**, 1895-8.
2. Caravella, J. & Lugovskoy, A. (2010). Design of next-generation protein therapeutics. *Curr Opin Chem Biol* **14**, 520-8.
3. Bolon, D. N., Wah, D. A., Hersch, G. L., Baker, T. A. & Sauer, R. T. (2004). Bivalent tethering of SspB to ClpXP is required for efficient substrate delivery: a protein-design study. *Mol Cell* **13**, 443-9.
4. Dreze, M., Charloteaux, B., Milstein, S., Vidalain, P. O., Yildirim, M. A., Zhong, Q., Svrzikapa, N., Romero, V., Laloux, G., Brasseur, R., Vandenhaute, J., Boxem, M., Cusick, M. E., Hill, D. E. & Vidal, M. (2009). 'Edgetic' perturbation of a C. elegans BCL2 ortholog. *Nat Methods* **6**, 843-9.
5. Bashor, C. J., Helman, N. C., Yan, S. & Lim, W. A. (2008). Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. *Science* **319**, 1539-43.
6. Kiel, C., Yus, E. & Serrano, L. (2010). Engineering signal transduction pathways. *Cell* **140**, 33-47.
7. Lippow, S. M. & Tidor, B. (2007). Progress in computational protein design. *Curr Opin Biotechnol* **18**, 305-11.
8. Karanicolas, J. & Kuhlman, B. (2009). Computational design of affinity and specificity at protein-protein interfaces. *Curr Opin Struct Biol* **19**, 458-63.
9. Mandell, D. J. & Kortemme, T. (2009). Computer-aided design of functional protein interactions. *Nat Chem Biol* **5**, 797-807.
10. Boas, F. E. & Harbury, P. B. (2007). Potential energy functions for protein design. *Curr Opin Struct Biol* **17**, 199-204.
11. Potapov, V., Cohen, M. & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel* **22**, 553-60.
12. Das, R. & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu Rev Biochem* **77**, 363-82.
13. Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M. & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043-54.
14. Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A. & MacBeath, G. (2008). Predicting PDZ domain-peptide interactions from primary sequences. *Nat Biotechnol* **26**, 1041-5.
15. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat Struct Biol* **9**, 553-8.
16. Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A. & Dahiyat, B. I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proc Natl Acad Sci U S A* **99**, 15926-31.

17. Treynor, T. P., Vizcarra, C. L., Nedelcu, D. & Mayo, S. L. (2007). Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc Natl Acad Sci U S A* **104**, 48-53.
18. Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci U S A* **107**, 19838-43.
19. Guntas, G., Purbeck, C. & Kuhlman, B. Engineering a protein-protein interface using a computationally designed library. *Proc Natl Acad Sci U S A* **107**, 19296-301.
20. Lippow, S. M., Moon, T. S., Basu, S., Yoon, S. H., Li, X., Chapman, B. A., Robison, K., Lipovsek, D. & Prather, K. L. Engineering enzyme specificity using computational design of a defined-sequence library. *Chem Biol* **17**, 1306-15.
21. Pantazes, R. J., Saraf, M. C. & Maranas, C. D. (2007). Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. *Protein Eng Des Sel* **20**, 361-73.
22. Parker, A. S., Griswold, K. E. & Bailey-Kellogg, C. Optimization of combinatorial mutagenesis. *J Comput Biol* **18**, 1743-56.
23. Wang, W. & Saven, J. G. (2002). Designing gene libraries from protein profiles for combinatorial protein experiments. *Nucleic Acids Res* **30**, e120.
24. Sammond, D. W., Eletr, Z. M., Purbeck, C. & Kuhlman, B. Computational design of second-site suppressor mutations at protein-protein interfaces. *Proteins* **78**, 1055-65.
25. Youle, R. J. & Strasser, A. (2008). The BCL-2 protein family: opposing activities that mediate cell death. *Nat Rev Mol Cell Biol* **9**, 47-59.
26. Chen, L., Willis, S. N., Wei, A., Smith, B. J., Fletcher, J. I., Hinds, M. G., Colman, P. M., Day, C. L., Adams, J. M. & Huang, D. C. (2005). Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol Cell* **17**, 393-403.
27. Gelinas, C. & White, E. (2005). BH3-only proteins in control: specificity regulates MCL-1 and BAK-mediated apoptosis. *Genes Dev* **19**, 1263-8.
28. Letai, A., Bassik, M. C., Walensky, L. D., Sorcinelli, M. D., Weiler, S. & Korsmeyer, S. J. (2002). Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer Cell* 2, 183-92.
29. Willis, S. N., Fletcher, J. I., Kaufmann, T., van Delft, M. F., Chen, L., Czabotar, P. E., Ierino, H., Lee, E. F., Fairlie, W. D., Bouillet, P., Strasser, A., Kluck, R. M., Adams, J. M. & Huang, D. C. (2007). Apoptosis initiated when BH3 ligands engage multiple Bcl-2 homologs, not Bax or Bak. *Science* 315, 856-9.
30. Leber, B., Lin, J. & Andrews, D. W. (2010) Still embedded together binding to membranes regulates Bcl-2 protein interactions. *Oncogene* **29**, 5221-30.
31. Feldhaus, M. J., Siegel, R. W., Opresko, L. K., Coleman, J. R., Feldhaus, J. M., Yeung, Y. A., Cochran, J. R., Heinzelman, P., Colby, D., Swers, J., Graff, C., Wiley, H. S. & Wittrup, K. D. (2003). Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat Biotechnol* **21**, 163-70.
32. Lee, E. F., Sadowsky, J. D., Smith, B. J., Czabotar, P. E., Peterson-Kaufman, K. J., Colman, P. M., Gellman, S. H. & Fairlie, W. D. (2009). High-resolution structural characterization of a helical alpha/beta-peptide foldamer bound to the anti-apoptotic protein Bcl-xL. *Angew Chem Int Ed Engl* **48**, 4318-22.
33. Lee, K. H., Han, W. D., Kim, K. J., Oh, B. H., unpublished.

34. Potapov, V., Reichmann, D., Abramovich, R., Filchtinski, D., Zohar, N., Ben Halevy, D., Edelman, M., Sobolev, V. & Schreiber, G. (2008). Computational redesign of a protein-protein interface for high affinity and binding specificity using modular architecture and naturally occurring template fragments. *J Mol Biol* **384**, 109-19.
35. Yosef, E., Politi, R., Choi, M. H. & Shifman, J. M. (2009). Computational design of calmodulin mutants with up to 900-fold increase in binding specificity. *J Mol Biol* **385**, 1470-80.
36. Guntas, G., Purbeck, C. & Kuhlman, B. Engineering a protein-protein interface using a computationally designed library. *Proc Natl Acad Sci U S A* **107**, 19296-301.
37. Grigoryan, G., Reinke, A. W. & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859-64.
38. Schreiber, G. & Keating, A. E. (2011). Protein binding specificity versus promiscuity. *Curr Opin Struct Biol* **21**, 50-61.
39. Smith, C. A. & Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol* **380**, 742-56.
40. Fisher, C. L. & Pei, G. K. (1997). Modification of a PCR-based site-directed mutagenesis method. *Biotechniques* **23**, 570-1, 574.
41. Hoover, D. M. & Lubkowski, J. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43.
42. Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M. & Wittrup, K. D. (2006). Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* **1**, 755-68.
43. Zhang, H., Nimmer, P., Rosenberg, S. H., Ng, S. C. & Joseph, M. (2002). Development of a high-throughput fluorescence polarization assay for Bcl-x(L). *Anal Biochem* **307**, 70-5.
44. Fu, X., Apgar, J. R. & Keating, A. E. (2007). Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-xL. *J Mol Biol* **371**, 1099-117.
45. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-90.

Chapter 4

Investigation and design of BH3 binding specificity toward different anti-apoptotic Bcl-2 proteins

Portions reprinted with permission of Elsevier B.V. from:

Dutta, S., Gulla, S., Chen, T. S., Fire, E., Grant, R. A. & Keating, A. E. (2010). Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. J Mol Biol 398, 747-62.

Portions of this chapter will be combined with other work done by Dr. Sanjib Dutta into a manuscript to be submitted later.

Collaborator notes:

Sanjib Dutta initiated the project, performed all yeast surface display experiments, and contributed a significant portion to all major sections in this chapter. Stefano Gulla and Emiko Fire performed the SPOT array experiments. Bob Grant assisted in solving the crystal structure of the complex between Mcl-1 and a Mcl-1 specific peptide.

Introduction

In the previous chapter I described how altering the sequence of the anti-apoptotic protein Bcl-xL affects its interaction specificity profile against different BH3 peptides. In this chapter, I investigate how sequence changes within the BH3 peptides influence their interaction specificity towards different anti-apoptotic proteins. Together, these studies contribute to on-going efforts in the Keating lab to broadly determine the factors that control binding specificity in the Bcl-2 family.

BH3-only proteins exhibit diverse binding specificities for anti-apoptotic Bcl-2 proteins. These are often measured using short peptides corresponding to the BH3 region of BH3-only proteins, for which the affinities of different anti-apoptotic proteins range over 10,000-fold. Most promiscuous are Bim and Puma, which bind to the five human anti-apoptotic proteins with dissociation constants in the low nanomolar range. In contrast, Bad and Noxa exhibit distinct preferences for some Bcl-2 proteins over others. Noxa derived peptides (denoted as Noxa-BH3) bind Mcl-1 and Bfl-1 with nanomolar affinity but show no detectable binding (> 100 μM) to other prosurvival family members. Conversely, Bad-BH3 binds with high affinity to Bcl-xL, Bcl-2, and Bcl-w but not to Mcl-1 or Bfl-1[2,3,4]. Mechanistically, selective binding profiles mean that only certain combinations of BH3-only proteins are able to kill cells[2]. The distinct binding characteristics of the prosurvival proteins are also relevant for small-molecule therapies that target them. ABT-737, the most effective known inhibitor, is selective for binding to Bcl-xL, Bcl-2, and Bcl-w[5] and has been shown to bind at the same site as the BH3 peptides[6]. However, cancers that rely on Mcl-1 to evade apoptosis are resistant to ABT-737 and related molecules[7].

Despite the importance of specificity in both the mechanism and the treatment of apoptotic misregulation in cancer, the sequence and structural determinants of binding specificity in Bcl-2

family members are still not completely understood. A number of studies have systematically addressed determinants of BH3 peptide binding to prosurvival Bcl-2 family members, and a few have addressed differential interactions with Bcl-xL versus Mcl-1[6,8,9]. Alanine and hydrophile scanning studies have been used to examine the effects of substitutions in several BH3 domains on binding to different anti-apoptotic proteins[6,8,9,10,11]. Strikingly, it has been demonstrated that Bim-BH3 variants with two or even three alanine mutations at conserved hydrophobic positions maintain high affinity for binding to Mcl-1 while losing binding affinity for Bcl-xL[9]. Guided by data generated from alanine and hydrophile scanning, Boersma et al.[8] combined pairs of point substitutions in Bim-BH3 to give peptides with nanomolar affinities for Mcl-1 that discriminated against Bcl-xL and vice versa. These mutants achieved >1000-fold specificity in the case of Mcl-1 binding and >100-fold specificity in the case of BclxL binding. These studies offered valuable insights into substitution effects in Bim-BH3.

In this chapter I describe studies using a position specific scoring matrix (PSSM) model directly derived from experimental SPOT array data[12] to study interaction specificity of BH3 peptides against different anti-apoptotic proteins. I first discuss how this simplified model, despite some apparent limitations, can provide insight into the sequence determinants of protein-protein interaction specificity. A significant amount of the analysis presented in this chapter is related to work done using yeast surface display to screen for novel BH3 sequences specific for Bcl-Xl over Mcl-1 or vice versa, and details of this can be found in Dutta et al[1]. Here, I describe my own efforts to develop and use PSSM models to design libraries of BH3 sequences to be screened for novel interaction specificity against different anti-apoptotic Bcl-2 proteins. I focus on how the limitations of the PSSM models affect decisions made for library design, and I

propose a design framework related to the one described in the previous chapter to cope with these limitations.

Results

SPOT array

Dr. Stefano Gulla carried out a substitution analysis of Bim-BH3 peptides in which 10 interface positions were mutated, one at a time, to all amino acids excluding Cys and Met (Fig. 4.1). SPOT arrays displaying 26-residue Bim-BH3 variants were constructed using solid-phase synthesis. Six hundred peptides were printed per membrane of 4 in. × 6 in., allowing the qualitative measurement of binding of hundreds of unique peptides simultaneously. Membranes of 200 spots each, including 170 Bim-BH3 variants, were probed with either 100 nM or 1 μM of Bcl-xL, Bcl-2, Bcl-w (only 100 nM), Mcl-1 and Bfl-1. The overall reproducibility of the data can be seen in the first column of each array, where every sequence is a repeat of the native. Good reproducibility was also observed for several mutant sequences that appeared two to three times on the membranes. Trends observed using 100 nM probe concentration were reproduced at the higher 1 μM concentration. For Bcl-xL and Mcl-1, additional interactions also become apparent at 1 μM, which is less true for Bfl-1 and Bcl-2. Additionally, Bcl-xL and Mcl-1 SPOT results agree qualitatively with previously reported binding studies for point mutations made in Bim-BH3 peptides and with a prior saturating substitution analysis at the 3a and 4a positions carried out using a phage ELISA technique[6]. No such comparison was made for the Bcl-2, Bcl-w and Bfl-1 SPOT results due to a relative lack of existing binding data between Bim-BH3 peptide mutants with these anti-apoptotic proteins in the literature.

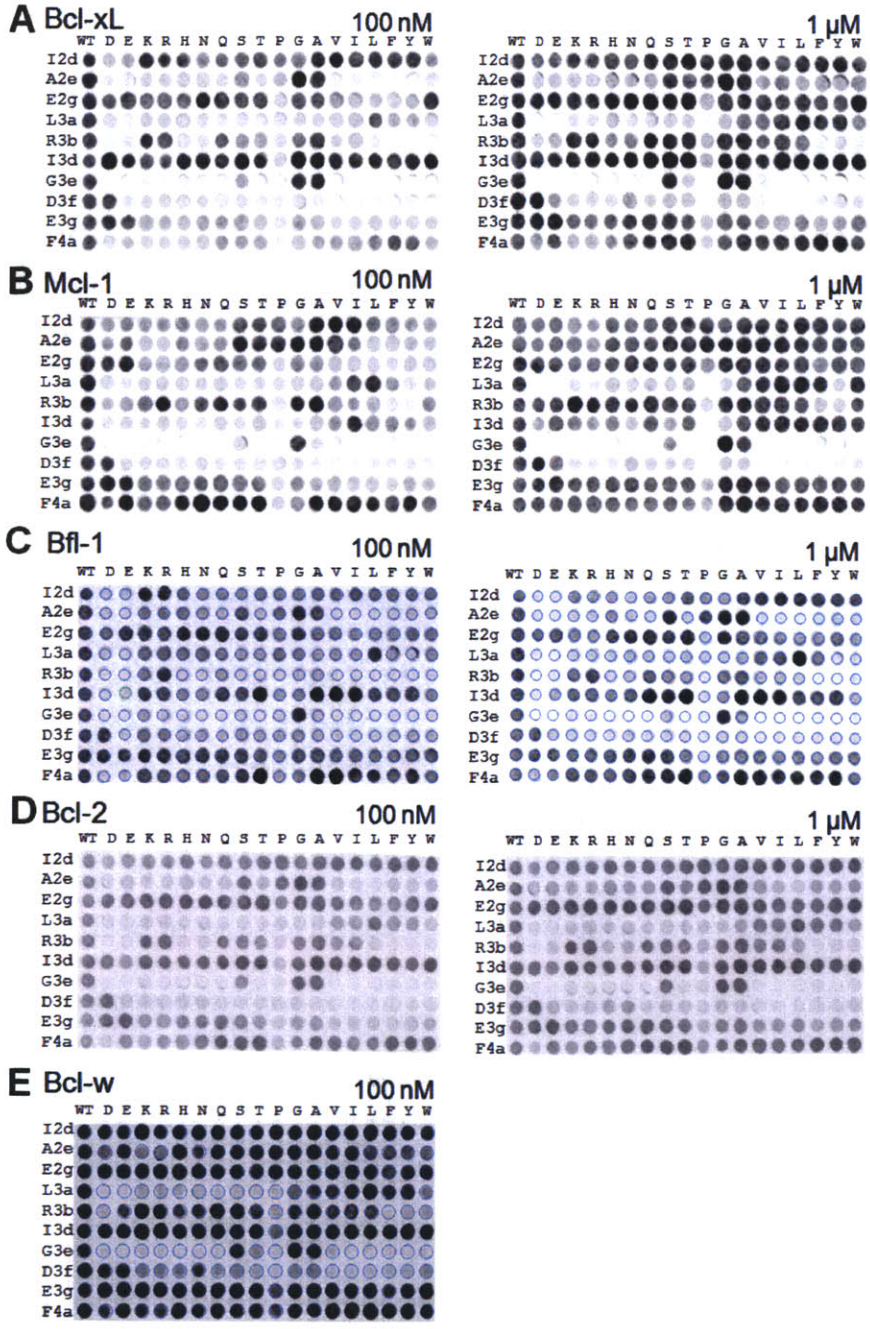


Figure 4.1 SPOT array substitution analysis of Bim-BH3 peptides binding to different anti-apoptotic proteins

(A) Bcl-xL, (B) Mcl-1, (C) Bfl-1, (D) Bcl-2, and (E) Bcl-w. The left and right panels used 100 nM and 1 μM of protein, respectively (except for Bcl-w, for which only a single experiment under 100 nM was performed). All spots in the leftmost column of each membrane show binding to the wild-type Bim-BH3 peptide. All other spots are point substitutions or a single repeat of the wild-type sequence in each row, with rows defining residue positions and columns indicating residue identities.

PSSM model building

Using SPOT data from the Bim-BH3 substitution analysis, we developed a position-specific scoring matrix (PSSM) to capture sequence features characteristic of binding to different anti-apoptotic proteins. We defined the score for amino acid i at position j binding to a specific anti-apoptotic protein R, $SR_{i,j}$, by taking the logarithm of the fluorescence intensity for the corresponding Bim point mutant normalized to the wild-type Bim intensity on the membrane. PSSM models were built for all 5 anti-apoptotic proteins, and only positions and amino acids covered by the SPOT analysis were included in the model. Different varieties of the PSSM models were also derived by averaging the scores obtained from SPOT membranes probed under different anti-apoptotic protein concentrations, or from additional experiments. Such models are described in more detail below.

Analysis of experimentally selected specific BH3 sequences using PSSM model

A previous study in our lab performed by Dr. Sanjib Dutta identified a diverse set of BH3 sequences that are specific for binding Bcl-xL over Mcl-1 and vice versa, as well as sequences that bind both Bcl-xL and Mcl-1 (Fig. 4.2, Table 4.1, 4.2, 4.3)[1]. The sequences were obtained from yeast surface display screening of a library of Bim-like BH3 sequences, with 6 positions being randomized. These 6 positions (2d, 3a, 3b, 3d, 3f, 4a) were a subset of the 10 positions probed on the SPOT array. These sequences provide an independent dataset to test our Bcl-xL and Mcl-1 PSSM models.

We used the PSSM to score each of the sequences isolated in yeast-display screening by summing score contributions from the six variable positions. As shown in Fig. 4.3E, this simple model does a good job separating sequences with different binding properties. Most of the Bcl-xL-specific sequences had high Bcl-xL scores and low Mcl-1 scores, whereas the Mcl-1-specific

sequences had low Bcl-xL scores and a range of Mcl-1 scores. Sequences of peptides that bound to both Mcl-1 and Bcl-xL generally had high Bcl-xL and Mcl-1 scores. Overall, the analysis shows that information about binding specificity for single-point mutants of Bim-BH3, as captured by the SPOT experiments, can be used to describe the specificities of the engineered sequences with a simple, linear model.

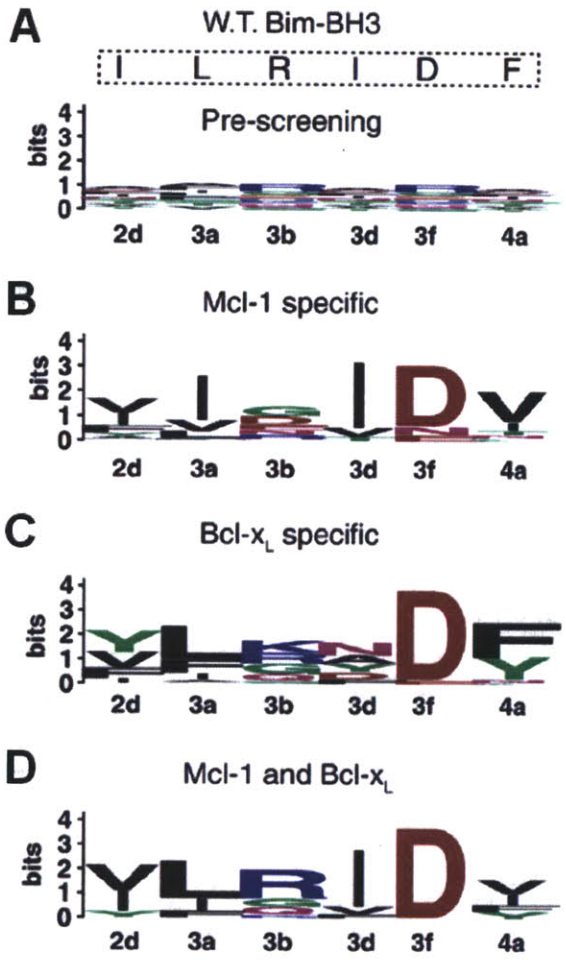


Figure 4.2 Sequence logos for sequences with different types of specificity identified from yeast surface display.
(a) The starting library prior to sorting with composition weighted by codon degeneracy; wild-type Bim residues are boxed at the top. (b) Mcl-1-specific peptides. (c) Bcl-xL-specific peptides. (f) Peptides that bound to both Bcl-xL and Mcl-1.

Table 4.1 Bcl-xL specific sequences identified from the yeast surface display screen

Clone	Position 2d	Position 3a	Position 3b	Position 3d	Position 3f	Position 4a
XD5	Y	L	R	N	D	F
XD6	Y	L	R	N	D	Y
XB9	Y	L	Q	N	D	F
XD9	Y	L	G	Y	D	Y
XG12	Y	L	G	A	D	F
XB6	Y	I	R	A	D	F
XC5	Y	I	G	Y	D	F
XD2	Y	I	Q	Y	D	F
XG10	Y	I	R	F	D	F
XD8	Y	V	K	Y	D	Y
XD1	Y	A	K	F	D	F
XB3	F	L	K	F	D	A
XF7	F	L	R	D	D	Y
XF8	F	L	K	N	D	Y
XG4	F	L	R	D	D	F
XF10	F	L	S	N	D	F
XB12	V	L	S	A	D	F
XE3	V	L	K	A	E	F
XH11	V	L	K	N	D	F
XA12	V	F	Q	N	D	F
XC4	V	F	K	F	D	Y
XF12	V	F	R	A	D	F
XH7	V	F	G	A	D	F
XE6	V	F	Q	A	D	Y
XC6	I	F	Q	A	D	Y
XG2	I	L	G	N	D	F
XB10	Y	L	K	D	D	F
XC9	Y	L	K	F	D	N
XA2	Y	L	K	Y	D	N
XD4	F	I	R	D	D	F
XA7	F	A	R	Y	D	F
XB4	F	L	S	Y	E	Y
XF9	V	L	G	D	D	F
XB8	V	L	G	N	D	Y
XC12	V	F	G	N	D	F
XE4	V	F	K	Y	D	T
XF1	I	F	Q	N	D	Y
XH8	I	L	Q	D	D	Y
XA1	Y	L	R	D	D	Y
XC10	Y	L	G	N	D	Y

Table 4.2 Mcl-1 specific sequences identified from the yeast surface display screen

Clone	Position 2d	Position 3a	Position 3b	Position 3d	Position 3f	Position 4a
MA5	F	V	N	I	D	V
MA9	F	V	G	I	D	V
MC10	F	V	G	I	D	I
MB2	F	I	D	I	D	V
MF12	F	I	E	I	D	V
MG1	F	F	S	I	D	V
MB1	I	I	D	I	D	V
MB9	I	I	G	I	D	T
MD5	I	I	G	I	N	V
MH1	I	I	G	T	D	V
MC3	I	I	D	I	N	I
ME9	I	I	N	V	D	I
MH11	I	I	D	I	D	T
MC11	I	I	D	I	D	F
MG2	I	I	R	I	E	Y
MA7	I	V	D	I	D	V
MA11	V	V	D	I	D	V
MG10	V	I	E	I	D	V
MD6	V	I	E	I	N	V
MG6	V	I	N	V	D	V
ME6	V	I	N	I	E	I
MA3	V	I	G	I	N	I
MB10	V	I	G	I	D	T
MH6	V	I	D	V	D	I
MD7	V	I	R	V	D	N
MH2	V	L	G	T	D	V
MF11	V	L	E	I	E	V
MA1	Y	V	Q	I	D	V
MA6	Y	L	E	I	D	V
MF2	Y	I	N	I	D	V
MH9	A	I	R	I	D	S
MB7	A	I	R	I	D	N
MB11	D	L	G	I	D	V

Table 4.3 Sequences binding both Bcl-xL and Mcl-1 identified from the yeast display screen

Clone	Position 2d	Position 3a	Position 3b	Position 3d	Position 3f	Position 4a
WT	I	L	R	I	D	F
PA3	V	L	Q	V	D	V
PA5	V	L	R	I	D	I
PB5	V	L	R	I	D	F
PB12	V	L	G	I	D	Y
PC5	I	L	R	F	D	I
PD2	V	L	K	I	D	Y
PD7	I	I	Q	V	D	I
PD11	Y	I	R	I	D	V
PE2	V	L	G	I	D	V
PE4	I	I	R	I	D	F
PE12	I	L	R	I	D	V
PF3	I	F	Q	I	D	I
PF6	I	F	R	I	D	V
PF8	V	L	G	I	D	V
PG4	V	L	R	I	D	Y
PH12	Y	I	R	I	D	I

Combinatorial library SPOT array

To explore sequence space more broadly, Dr. Sanjib Dutta synthesized combinatorial library SPOT arrays. We identified residues that occurred with high frequency in selected sequences from the yeast-display screening in our previous study: Ile (wild type), Ala, and Phe at position 2d; Leu (wild type), Ile, Phe, and Ala at position 3a; Arg (wild type) and Asp at position 3b; Ile (wild type), Phe, Asp, Asn, and Ala at position 3d; and Phe (wild type), Val, and Asn at position 4a. From this reduced library, we synthesized all 360 possible sequences. The resulting membranes, referred to here as library arrays, were probed with 100 nM Mcl-1 or Bcl-xL. Some interactions of interest are shown in Fig. 4.3A and 4.3B, and the whole library array can be found in Dutta et al.[1]. The library arrays included a wider range of sequence contexts and highlighted

specificity-determining residues not evident in the Bim-BH3 substitution arrays. This was valuable for model building and interpretation (see below).

Further improvement of the PSSM model for Bcl-xL and Mcl-1

The initial PSSM model performed well, and we explored simple ways in which it could be improved. Although we currently lack the large amount of quantitative data required to describe synergy between peptide positions, even simple PSSM models can potentially be improved by obtaining better estimates of single-position effects. Therefore, we used data from the library arrays to construct a second PSSM, which allowed us to derive mutational scores averaged over multiple contexts for some key substitutions. Evaluating substitutions in multiple contexts also provided a larger dynamic range for the assay. Using the revised PSSM model, we obtained better separation of scores on the Mcl-1 binding axis (Fig. 4.3F). Notably, the percentage of Mcl-1-specific peptides having Mcl-1 scores higher than the highest-scoring Bcl-xL-specific peptide along this axis increased from 33% to 85%. Much of this change was attributable to a significantly more favorable score for Val at 4a binding to Mcl-1, when averaged over the library SPOT sequences. Although this was not obvious from our single-substitution SPOT arrays (Fig. 4.1B), sequences with Val at 4a exhibited significantly enhanced binding to Mcl-1 compared with the wild-type residue Phe in the context of destabilizing mutations at other positions (e.g., Phe, Asp, Asn, or Ala at position 3d or Asp at position 3b) (Fig. 4.3A and 4.3B). Competition binding assays confirmed that a Phe4aVal mutation in Bim-BH3 increased affinity for Mcl-1 by more than 10-fold (Ki of ~100 pM) while reducing affinity for Bcl-xL by ~30-fold (Ki of ~30 nM) (Fig. 4.3C and 4.3D)

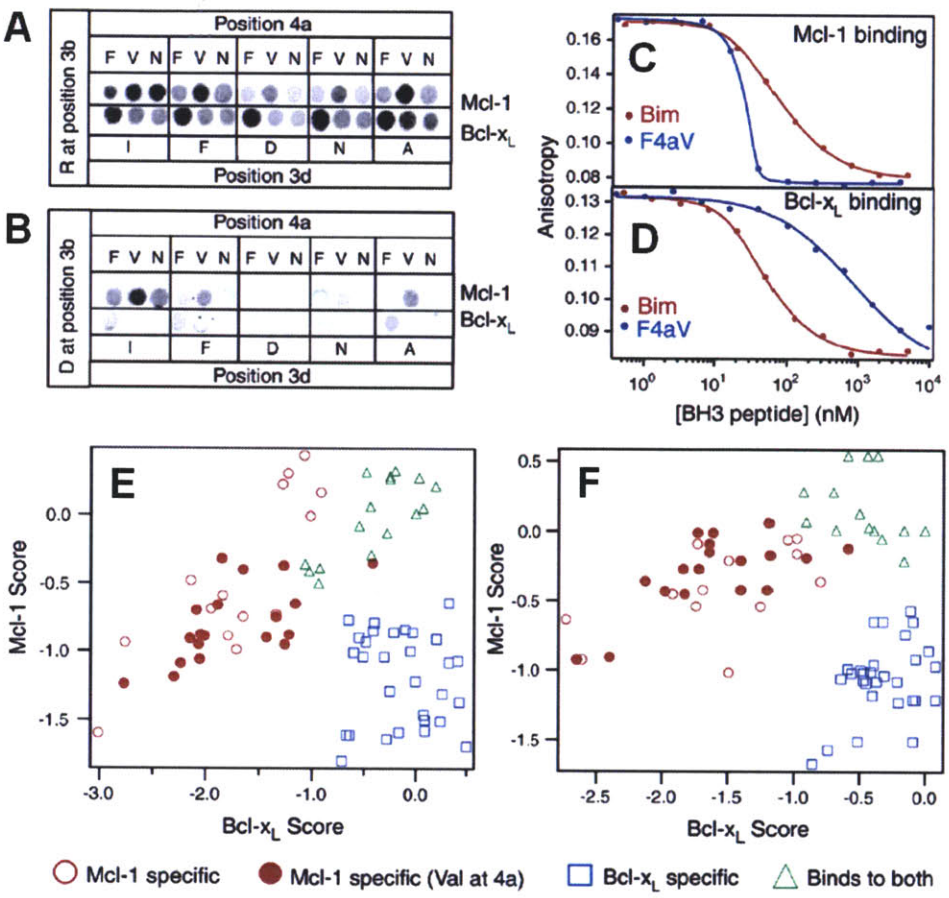


Figure 4.3 A model built using the SPOT array data captures the specificities of sequences identified using yeast display.

(A and B) A section of the library arrays showing position 4a substitutions. (A) Each boxed set of three spots shows substitution at position 4a with Phe, Val or Asn. Mutations were made with different residues at position 3d, as indicated, with all other residues identical to wild-type Bim-BH3. SPOTS in the top or bottom rows were probed with 100 nM Mcl-1 or 100 nM Bcl-x_L, respectively. (B) Same as (A) but for mutations made in the context of Asp at 3b. (C) Effect of a Phe-to-Val substitution at position 4a in Bim-BH3 on binding to Mcl-1 (C) or Bcl-x_L (D) in fluorescence competition binding assays as described in Dutta et al[1]. (E) Engineered BH3 peptide sequences from the yeast screen were scored using a PSSM based on the Bim-BH3 substitution array data. The points plotted correspond to: Mcl-1 specific peptides (red circles), Mcl-1 specific peptides with Val at position 4a (red filled circles); Bcl-x_L specific peptides (blue squares), peptides that bound to both proteins (green triangles). (F) The same plot constructed with a PSSM that included the SPOT library array data; this model gave better separation of Mcl-1 binders vs. non-binders along the Mcl-1 score axis.

Bfl-1 library design

The goal for this part of the study was to design BH3 libraries to be screened by yeast surface display for sequences that are specific for one anti-apoptotic protein over all others. We started by designing a library to be screened for sequences showing binding specificity for Bfl-1 over Bcl-xL, Mcl-1, Bcl-w and Bcl-2, guided by the PSSM models derived from SPOT array results for the 5 anti-apoptotic proteins. For Bcl-xL and Mcl-1, we used the revised/improved PSSM model as described in the previous section. For Bfl-1 and Bcl-2, PSSM scores were obtained by averaging the scores derived from SPOT membranes probed under 100 nM and 1 μM concentration of the anti-apoptotic proteins. For Bcl-w, the PSSM scores were derived from the membrane probed with a concentration of 100 nM Bcl-w. We then defined two classes of residues, non-disruptive and specific, at each designed position according to their PSSM scores. A residue was defined as non-disruptive if its Bfl-1 PSSM score was among the top 50% of the Bfl-1 scores for all amino acids across all position on the membrane. A non-disruptive residue was further classified as specific if the difference of its Bfl-1 PSSM score with the PSSM score for another anti-apoptotic protein was greater than $\log_{10}(1.5)$ and ranked among the top 33% of all such differences for all amino acids across all positions. Four types of specificity residues (Bfl-1 over Bcl-xL, Bfl-1 over Mcl-1, Bfl-1 over Bcl-2 and Bfl-1 over Bcl-w) were defined accordingly.

Next we chose degenerate codons at each designed positions to enrich the library with the non-disruptive and specific residues defined above. We wanted to enrich the library with two types of combinatorial diversities: (1) Diversity from all designed positions occupied by non-disruptive or native residues and (2) Diversity from all designed positions occupied by specific or native residues. We formulated an optimization procedure based on integer linear

programming (ILP) to maximize the product of the number of these two combinatorial diversities under a constraint of library size of 10^7, a conservative number for adequate experimental coverage of the library when doing yeast surface display. The optimization also favored selection of degenerate codons encoding all types of specific residues available at a designed position, and ones encoding amino acids with greater chemical diversity. A more detailed description of the optimization procedure can be found in Materials and Methods. The resulting optimized library is shown in Table 4.4.

Table 4.4 Bfl-1 library design results

	Non-disruptive residues	vs. Bcl-xL	vs. Mcl-1	vs. Bcl-2	vs. Bcl-w	residues encoded
I2d	ACFGHIKMLPRTV WY	-	WKY	-	-	FIKLMNYZ (WWK)
A2e	ACGHPS	HS		H	-	ADHPSY (BMT)
E2g	ACDEFGHIKNQRST WY	Y	FGHIKV Y	-	GT	CDEFGIKLMNRSVWYZ (DDK)
L3a	FILMNV	NV	N	N	N	DHILNV (VWC)
R3b	AKQR	-	-	-	-	Not randomized
I3d	ACFGHIMKLNQRST VY	-	AFGHKN QRSTVY	-	-	ACDFGHILNPRSTVY (NNT)
G3e	G	-	-	-	-	Not randomized
D3f	D	-	-	-	-	Not randomized
E3g	ACDEFGHIKLMNQ RSTVWY	AFHIKLN RQSTVW Y	AFILWV Y	FIKLRV WY	-	AEIKLPQTV (VHA)
F4a	ACFGHIKLMNQRST VWY	AGHIKL QRSTW	-	K	KR	CFIKLMNRSWYZ (WDK)

For each position, non-disruptive residues (second column) and residues predicted by the PSSM models to favor binding to Bfl-1 over the indicated anti-apoptotic protein (the next four columns) are listed. The final column shows residues included in the designed library (encoded by the degenerate codon in parentheses), as optimized using the ILP framework.

Bcl-xL/Bcl-2/Bcl-w library design

To identify sequences specific for each of Bcl-xL, Bcl-2 and Bcl-w, I designed a single joint library instead of 3 separate libraries. This is because the Bcl-xL, Bcl-2 and Bcl-w sequences are highly similar to one another within the interface region. The idea was to enrich a library in sequences that would bind these three receptors in preference to Mcl-1 and Bfl-1, and also to include sequence elements predicted to favor Bcl-xL, Bcl-2 or Bcl-w individually. Predicted non-disruptive residues for this library were predicted based on the Bcl-xL PSSM. At all positions, we approximated the predicted specificity of Bcl-xL/Bcl-w/Bcl-w over Mcl-1 and Bfl-1 as the PSSM predicted specificity of Bcl-xL over Mcl-1 and Bfl-1. Specificity among Bcl-xL/Bcl-w/Bcl-2 was predicted only for Bim BH3 positions 2g, 3a, 3d, and 3g. These 4 positions were selected because they contact positions occupied by different amino acids among Bcl-xL/Bcl-w/Bcl-2 when aligned to the Bcl-xL/Bim complex crystal structure (PDB ID 3FDL). The design procedure was otherwise similar to that for Bfl-1 library design, and is described in more detail in Materials and Methods. The resulting optimized library is shown in Table 4.5.

Table 4.5 Bcl-xL/Bcl-2/Bcl-w library design results

	Non-disruptive residues	vs. Mcl-1	vs. Bfl-1	intra	residues encoded
I2d	ACEFGHIKLMNPQR STVWY	HKQRY	EHS	-	CDEFGHIKLMNQRSVW YZ (NDK)
A2e	ACGPS	-	-	-	A
E2g	ACDEGHIKLMNQR STVWY	AHKNR WV	W	AFGILT VWY	CDEFGIKLMNRSVWYZ (DDK)
L3a	AFILMY	Y	Y	AY	ADFHLPSVY (BHC)
R3b	ACGKQRST	-	AK	IS	AEGIKLRSTVZ (DNA)
I3d	ACDEFGHIKLMNQ RSTVWY	ADEFGH KLNQRS T VWY	DEGHN W	E	ACDEFGIKLMNRSTVW YZ (DNS)
G3e	ACGS	A	A	-	AG (GSC)
D3f	D	-	-	-	Not randomized
E3g	ACDEGHKNQRS	-	-	-	EGQR (SRA)
F4a	AFILMNQSTVWY	-	-	-	Not randomized

For each position, non-disruptive residues (second column) and residues predicted by the PSSM models to favor binding to Bcl-xL over the indicated anti-apoptotic protein (the next two columns) are listed. Listed under the column "intra" are residues that are either "Bcl-xL over Bcl-2", "Bcl-xL over Bcl-w", "Bcl-2 over Bcl-xL", "Bcl-2 over Bcl-w", "Bcl-w over Bcl-xL", or "Bcl-w over Bcl-2" specific. The final column shows residues included in the designed library (encoded by the degenerate codon in parentheses), as optimized using the ILP framework.

Screening

Screening of both libraries is being conducted by Dr. Sanjib Dutta using yeast-surface display techniques, as in Dutta et al. Successive rounds of selection for binding to the desired target in the presence of competitors have enriched certain sequences. Analysis of these, and of the success of the design strategy, will be included in a future joint publication.

Discussion

The PSSM model introduced in this chapter is arguably one of the simplest experimentally parameterized models. The ability to probe interaction between more than 200 mutant peptides and an anti-apoptotic protein in a single experiment provides a cost-effective, facile way to map the peptide sequence space comprehensively. However, there are many limitations with such models, arising from the SPOT experiment itself and the formulation. First of all, neither the SPOT experiment nor the model derived takes into account any interactions among different BH3 peptide positions. Secondly, the SPOT results are semi-quantitative at best, due to issues such as variation of the peptide synthesis yield on the membrane, the dynamic range of the signal, and non-equilibrium conditions. Below I suggest how, despite these caveats, it is still possible to obtain useful insights into the relation between sequence and interaction specificity, especially when combined with other experimental data, including the yeast display selected sequences (Table 4.1, 4.2, 4.3) and the library SPOT arrays described in Results. I focus in the analysis below on the PSSM models for Bcl-xL and Mcl-1, as these are the two anti-apoptotic proteins for which we have the appropriate data for analysis. In the end we shift the discussion to library design based on such models.

Analysis of experimentally selected sequences using the PSSM model

We used the substitution arrays to construct a PSSM and showed that this model can separate Mcl-1-specific sequences from Bcl-xL-specific sequences and from sequences of peptides that bind with high affinity to both receptors (Fig. 4.3E). Thus, although we cannot rule out synergistic effects between positions in Bim-BH3 that may influence binding, much of the specificity observed in the sequences identified from yeast-display screening can be explained by

a simple, linear, and additive model. Importantly, this model was derived independent of knowledge of these sequences.

To see if the Bim-BH3-based PSSM could be improved, and to explore the effects of point mutations in the context of sequences selected from the yeast-display library rather than Bim-BH3, we used the library arrays (Fig. 4.3A and 4.3B). The PSSM model built using data from the library arrays was similar to that based on the Bim-BH3 substitution analysis, but it did a better job of discriminating high-affinity versus low-affinity binding to Mcl-1 (Fig. 4.3F). We traced this effect largely to the role of stabilizing mutations at position 4a and confirmed using solution binding studies that Val at this site is stabilizing relative to wild-type Phe for Mcl-1 binding (Fig. 4.3C and 4.3D).

The two PSSM models differed in two ways: First, the library arrays allowed us to evaluate the effects of key point substitutions using average values collected over many Bim-like sequences. These averages may provide better estimates of the influence of mutations in the engineered peptides, and the larger numbers of measurements also make them less sensitive to noise. Second, the high affinity of native Bim-BH3 for Mcl-1 and Bcl-xL saturates the signal in the SPOT arrays for many sequences and thus masks the effects of stabilizing mutations. Because of this, the Bim-BH3 substitution array matrix incorrectly assigned similar weights to Val and Phe at position 4a for Mcl-1 binding. Our work indicates that both the concentration used for the SPOT experiments (compare the left and right panels for Fig. 4.1A and 4.1B) and the sequence context in which mutations are made (Fig. 4.3A and 4.3B) can be important for providing appropriate mutational data to parameterize a predictive model.

Mechanism for Bcl-xL vs. Mcl-1 specificity

Using the SPOT data as a guide, we investigated the mechanisms used to establish interaction specificity in the peptides identified by the previous yeast surface display screen referred to in this study. We defined three classes of substitutions according to interaction weights from the arrays (Table 4.6). Class 1 and class 2 substitutions were specific for one anti-apoptotic protein over another. The difference between these two classes is that class 1 substitutions retained strong binding to the desired target on the arrays, whereas class 2 substitutions achieved specificity at the expense of some stability. Class 3 substitutions were highly destabilizing for binding to both anti-apoptotic proteins, without any discernable preference. Interestingly, most of the substitutions identified as class 1 based on the arrays were highly represented in the specific sequences identified by yeast display screening.

Many class 1 substitutions occurred in position 3d or 4a. At position 3d, both Mcl-1-specific sequences and sequences of peptides that bound both anti-apoptotic proteins were largely constrained to the wild-type Bim residue Ile (Fig 4.2B). In contrast, sequences specific for Bcl-xL spanned a range of residues, including polar residues, but never Ile (Fig. 4.2C). In co-crystal structures of Bim in complex with Bcl-xL versus Mcl-1, the 3d site is less tightly packed in Bcl-xL, where it is located next to a less helical α2/α3 region of the receptor; this may explain the observed permissiveness[13]. Thus, the class 1 mutations favoring Bcl-xL at 3d (Ala, Asp, Asn, Phe, Tyr, Thr) appear to be key specificity determining factors disfavoring Mcl-1 binding.

Table 4.6 Classification of representative substitutions observed in selected sequences according to their intensities as measured on the substitution SPOT array

Position	Substitutions	Class	Specificity
2d	F/Y	1[a]	Bcl-xL
3a	I	1[a]	Mcl-1
	A	3[c]	-
3b	N	1[a]	Mcl-1
	D,E	2[b]	Mcl-1
3d	A/D/N/F/Y/T/V	1[a]	Bcl-xL
3f	E/N	3[c]	-
4a	N/S/V/T/I	1[a]	Mcl-1

[a] Normalized signal intensity for binding to one anti-apoptotic protein more than 2-fold of that of another. Signal intensity for the preferred receptor, ≥0.7.
[b] Normalized signal intensity for binding to one anti-apoptotic protein more than 2-fold of that of another. Signal intensity for the preferred anti-apoptotic protein, ~0.2–0.3.
[c] Normalized signal intensity for both anti-apoptotic proteins, <0.2.

At position 4a, the sequence logos in Fig. 4.2C emphasize that Bcl-xL is selective for large aromatics, while Mcl-1 can accommodate multiple substitutions (Fig. 4.2B), with Asn, Ser, Val, Thr, and Ile assigned as class 1 mutations favoring Mcl-1 binding. The co-crystal structure of Mcl-1 with one of the Mcl-1 specific peptides, MB7, shows that Asn can be easily accommodated at position 4a[1], without any significant local perturbation, in agreement with previous observations that this site is more open and solvent-exposed in Mcl-1 compared with Bcl-xL[6,8,9,14].

At position 2d, two class 1 mutations favoring Bcl-xL (Phe and Tyr) were very common in Bcl-xL-specific sequences (Fig. 4.2C). It is interesting that the BH3 region of Bad, which is highly specific for Bcl-xL over Mcl-1, also has a Tyr at the same position. Mutational studies in Bad have confirmed that this residue influences binding specificity[15]. Ile at 3a is a class 1 substitution for Mcl-1, and this is prominent in the Mcl-1-specific sequence logo.

For position 3b, the sequence logo reveals relatively low information. However, the substitutions Asn (class 1 for Mcl-1) and Glu or Asp (class 2 for Mcl-1) are present in the Mcl-1-specific sequences and completely absent from the Bcl-xL specific sequences (Table 4.1 and 4.2).

An examination of individual sequences identified in the yeast screen shows that all contain more than one substitution from wild-type Bim-BH3. Most Bcl-xL and some Mcl-1-specific peptides combined multiple class 1 mutations, including Bcl-xL-specific peptide XD5 (two class 1 substitutions: Tyr at position 2d and Asn at position 3d) and Mcl-1-specific MB9 (two class 1 substitutions: Ile at position 3a and Thr at position 4a) (Table 4.1 and 4.2). Interestingly, many Mcl-1-specific sequences combined class 1 with class 2/3 substitutions (such as Asp/Glu at position 3b or Asn/Glu at position 3f), thereby achieving specificity but sacrificing stability. Many of these sequences also included Val/Ile at position 4a as the class 1 mutation. Therefore, we speculated that Val/Ile, in addition to providing specificity as class 1 substitutions, might provide stability to compensate for destabilizing mutations. Interestingly, as shown in Fig. 4.3, the point mutation Phe4aVal in Bim-BH3 increased Mcl-1 binding affinity and conferred a significant preference for binding Mcl-1 over Bcl-xL. This type of single amino-acid substitution would be missed in the screen, which eliminated all clones that bound Bcl- xL at 1 μM concentration[1]. These observations point to an interesting strategy to satisfy the requirements of the screen—that is, combining substitutions that destabilize binding for both receptors (to meet the specificity constraint) with ones that selectively enhance binding for the receptor of interest (to meet the stability constraint). Using the above analysis, we could rationalize the sequence patterns for most of the specific sequences.

We would like to emphasize that the classifications and interpretations presented above are based largely on SPOT experiments but not more rigorous quantitative measurements of binding

affinity. Therefore, we avoid some of the more subtle issues, such as the role of substitutions that are not clear-cut in our classification scheme, and questions about whether multiple specificity determinants are synergistic or simply additive. Despite these simplifications, we show that a framework based on a simple SPOT/PSSM analysis can logically explain many sequence–function relationships that underlie the observed behavior of the specific peptides. Whereas our model is imperfect and leaves the detailed behavior of various examples unexplained, the power of experimental screening has nevertheless provided sequences that combine different substitutions to achieve multiple objectives, whether these combinations follow our intuition or have more subtle effects.

Library design

In this section I discuss the rationale behind library design guided by PSSM. I focus the discussion on Bfl-1 library design, although the same general concepts stand for the Bcl-xL/Bcl-2/Bcl-w library design as well.

As described before, despite evidence that SPOT PSSM models can provide insight into the sequence determinants of interaction specificity, one should not over-estimate its predictive power. One bottleneck preventing the SPOT PSSM models from making meaningful specificity predictions stems from the limited dynamic range of SPOT signals. For example, it is difficult to evaluate whether residues with strong or saturated SPOT signals enhance or weaken binding relative to the native residue, and hence their roles in specificity. A larger number of such residues were present on the Bcl-w, Bcl-2, and Bfl-1 membrane compared to those on Bcl-xL and Mcl-1. This could obviously reflect the possibility that Bcl-w, Bcl-2 and Bfl-1 are on average more tolerant of interaction with Bim BH3 point mutants. However, it could also be

explained by different anti-apoptotic proteins simply possessing different dynamic ranges for the SPOT signals.

Regardless of the origin, the observations noted above could complicate prediction of global specificity for binding Bfl-1 in preference to all other 4 anti-apoptotic proteins. For example, only a few residues were predicted to be Bfl-1 over Bcl-w specific, and it could be risky to enforce inclusion of these residues in all library members without further verification of their influence on binding. We therefore identified and considered all residues predicted to confer specificity against any of the four off-target receptors. Global specificity was addressed by requiring that the degenerate codons chosen span residues predicted to confer specificity against each receptor when applicable (see Materials and Methods). We reasoned that this should allow a diverse number of ways to combine residues of different specificity types from each designed position in the library, and hopefully the screening could identify the correct combinations of the ones behaving as predicted. The extent of inclusion of these residues was influenced by the library size constraint, and also by the amino acid diversity of the degenerate codons available. At the same time, even if no predictions were correct for one or more particular binary specificity class (e.g. Bfl-1 over Bcl-w), as long as other specificity predictions were accurate to some extent, the library would still be sampling a sequence space guided in a useful manner. A strategy emphasizing diversity can also be rewarded by capturing important residues missed by predictions.

In addition to enriching residues predicted to be specific according to SPOT PSSMs, we also aimed to enrich residues showing at least moderate binding on the Bfl-1 SPOT membrane (the non-disruptive residues). Note that because all predicted specificity residues were required to have a specified minimal Bfl-1 SPOT signal, their inclusion was given priority in our library

design scheme and these residues indeed dominated the designed libraries in this study (Table 1, 2). However, one could imagine, in cases where fewer specificity residues were predicted, a more inclusive strategy could be rewarded for the same reasoning described above. Again the extent of inclusion of such residues would be constrained by other factors such as the library size.

Materials and Methods

PSSM model

The original PSSM score for amino acid *i* at position *j* binding to a specific anti-apoptotic protein R, $SR_{i,j}$, was obtained by taking the logarithm ($\log_{10}$) of the ratio of the fluorescence intensity for the corresponding Bim-BH3 point mutant to the intensity of wild-type Bim-BH3 (averaged over all wild-type spots) on the membrane.

To derive the revised PSSM models for Bcl-xL and Mcl-1, we used spots from the library array that had raw signals $>10^{6.5}$. To score the contribution of residue *i* at position *j*, we computed the log of the ratio of the average signal for peptides with residue *i* at position *j* to the average signal for peptides with the native Bim-BH3 residue at position *j*. For residues not included in the library arrays, the substitution array data were used. However, scores for residues that had normalized intensities greater than 1 in the substitution arrays were reduced to 1. In addition, the score for Ile at 4a was assigned the same value as Val at 4a because Ile was not included in the library SPOTS and Val and Ile had similar scores from the substitution arrays.

For library design, we used the revised model for Bcl-xL and Mcl-1. For Bcl-2 and Bfl-1, the models were obtained by averaging the PSSM scores derived from membranes probed with 100 nM and 1 μM of anti-apoptotic proteins. For Bcl-w, the model was derived from the membrane probed with 100 nM Bcl-w. In all models used for design, normalized intensities for all residues

with values greater than 1 were capped at 1 before use in deriving the PSSM score (except for the revised models as described above). As Met and Cys substitutions were not printed on the membrane, their scores were defined as those of Leu and Ser respectively when predicting "non-disruptive" residues. However, they were not scored for the specificity predictions.

Bfl-1 library design

The definition of non-disruptive residues and 4 different types of specific residues (Bfl-1 over Bcl-xL, Mcl-1, Bcl-w, or Bcl-2) is given in the Results. In addition, Pro was removed from consideration as a potential specific residue. Four quantities were defined for each degenerate codon j at position i: (1) the size, s_{ij}, which is the number of unique tri-nucleotides within the codon, (2) nd_{ij}, the number of non-disruptive residues encoded by the codon, (3) sp_{ij}, the number of specific residues (considering all 4 different types) encoded by the codon, and (4) m_{ij}, the number of "misses" in chemical diversity for the codon. The metric m_{ij} was defined for codons at positions 2d, 2g, 2a, 3b, 3a, 3d, 4a. Amino acids were divided into different classes according to their physicochemical properties, and then the number of classes with no representation from the amino acids encoded by the codon was counted. For the more buried positions 2d, 3a, 3d, 4a, the classes were [A], [L], [IV], [FY]. For the more exposed positions 2g, 3b, 3g, these classes were [AG], [DE], [KR], [NQ], [ST]. A "miss" was scored for a class only if at least one amino acid from that class was designated non-disruptive. A codon was considered more chemically diverse if it has a lower m_{ij}.

At each designed position, we only considered degenerate codons that encode (1) the native Bim BH3 amino acid and (2) at least one of each type of specific residue present at that position. The set of remaining degenerate codons was trimmed further by comparing every two. If a

degenerate codon had a larger s_{ij}, a smaller sp_{ij}, and a larger m_{ij} than another codon, then the first codon was considered dominated by the second and was eliminated from the pool. The elimination process was repeated until no degenerate codon dominated any other codon in the remaining set. Optimization of degenerate codon combinations, out of the remaining pool of codons J_i at each designed position i, was performed by solving the following integer linear programming problem:

$$\text{Max } \sum_i \sum_{j \epsilon Ji} c_{ij} \log(nd_{ij}) + \sum_i \sum_{j \epsilon Ji} c_{ij} \log(sp_{ij})$$

$$\text{subject to } \sum_i \sum_{j \epsilon Ji} c_{ij} \log(s_{ij}) <= 7$$

$$\text{subject to } \sum_i \sum_{j \epsilon Ji} c_{ij} m_{ij} <= 4$$

$$\text{subject to } \sum_{j \epsilon Ji} c_{ij} = 1 \text{ for each position } i$$

Where $c_{ij} = 1$ if codon j was picked at position i, and 0 otherwise. For the winner codon j picked at each position i, $\sum_i \log(nd_{ij}) = \log(\prod_i nd_{ij})$ is the logarithm of the number of unique protein sequences encoded with all designed positions occupied by non-disruptive residues, $\sum_i \log(sp_{ij}) = \log(\prod_i sp_{ij})$ is the logarithm of the number of unique protein sequences encoded with all designed positions occupied by specific residues, and $\sum_i \log(s_{ij}) = \log(\prod_i s_{ij})$ is the library size (or the number of unique DNA sequences in the library) as described in the text. $\sum_i m_{ij}$ is the total number of misses in chemical diversity across all positions and we manually picked 4 as a threshold. The problem was solved using the glpsol solver in the GLPK package (GNU MathProg).

Bcl-xL/Bcl-2/Bcl-w library design

Most of the design procedure is the same as that for Bfl-1 library design, with a few modifications. "Non-disruptive" residues were defined based on the Bcl-xL PSSM. As describe in Results, for positions 2d, 2e, 3b, 3e, 3f and 4a, only two types of specificity residues were predicted, "Bcl-xL over Mcl-1" and "Bcl-xL over Bfl-1". For positions 2g, 3a, 3d and 3g, 6 more types of specificity residues, "Bcl-xL over Bcl-2", "Bcl-xL over Bcl-w", "Bcl-2 over Bcl-xL", "Bcl-2 over Bcl-w", "Bcl-w over Bcl-xL", and "Bcl-w over Bcl-2" were predicted. As for Bfl-1 library design, only codons that encode (1) the native Bim BH3 amino acid and (2) all different types of specific residues present at each design position were considered.

Acknowlegements

We thank E. Genillo, J.A. Tan, W. Garcia-Beltran for assistance, and R. Cook and the M.I.T Biopolymers facility for peptide and SPOT array synthesis. We thank members of the Keating lab for helpful discussions and the Baker and the Bell labs for use of equipment. This work was funded by NIGMS awards GM084181 and P50-GM68762.

References

1. Dutta, S., Gulla, S., Chen, T. S., Fire, E., Grant, R. A. & Keating, A. E. (2010). Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. J Mol Biol 398, 747-62.
2. Chen, L., Willis, S. N., Wei, A., Smith, B. J., Fletcher, J. I., Hinds, M. G., Colman, P. M., Day, C. L., Adams, J. M. & Huang, D. C. (2005). Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol Cell* 17, 393-403.
3. Certo, M., Del Gaizo Moore, V., Nishino, M., Wei, G., Korsmeyer, S., Armstrong, S. A. & Letai, A. (2006). Mitochondria primed by death signals determine cellular addiction to antiapoptotic BCL-2 family members. *Cancer Cell* 9, 351-65.
4. Kuwana, T., Bouchier-Hayes, L., Chipuk, J. E., Bonzon, C., Sullivan, B. A., Green, D. R. & Newmeyer, D. D. (2005). BH3 domains of BH3-only proteins differentially regulate Bax-mediated mitochondrial membrane permeabilization both directly and indirectly. *Mol Cell* 17, 525-35.
5. Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., Bruncko, M., Deckwerth, T. L., Dinges, J., Hajduk, P. J., Joseph, M. K., Kitada, S., Korsmeyer, S. J., Kunzer, A. R., Letai, A., Li, C., Mitten, M. J., Nettesheim, D. G., Ng, S., Nimmer, P. M., O'Connor, J. M., Oleksijew, A., Petros, A. M., Reed, J. C., Shen, W., Tahir, S. K., Thompson, C. B., Tomaselli, K. J., Wang, B., Wendt, M. D., Zhang, H., Fesik, S. W. & Rosenberg, S. H. (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* 435, 677-81.
6. Lee, E. F., Czabotar, P. E., Smith, B. J., Deshayes, K., Zobel, K., Colman, P. M. & Fairlie, W. D. (2007). Crystal structure of ABT-737 complexed with Bcl-xL: implications for selectivity of antagonists of the Bcl-2 family. *Cell Death Differ* 14, 1711-3.
7. van Delft, M. F., Wei, A. H., Mason, K. D., Vandenberg, C. J., Chen, L., Czabotar, P. E., Willis, S. N., Scott, C. L., Day, C. L., Cory, S., Adams, J. M., Roberts, A. W. & Huang, D. C. (2006). The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. *Cancer Cell* 10, 389-99.
8. Boersma, M. D., Sadowsky, J. D., Tomita, Y. A. & Gellman, S. H. (2008). Hydrophile scanning as a complement to alanine scanning for exploring and manipulating protein-protein recognition: application to the Bim BH3 domain. *Protein Sci* 17, 1232-40.
9. Lee, E. F., Czabotar, P. E., van Delft, M. F., Michalak, E. M., Boyle, M. J., Willis, S. N., Puthalakath, H., Bouillet, P., Colman, P. M., Huang, D. C. & Fairlie, W. D. (2008). A novel BH3 ligand that selectively targets Mcl-1 reveals that apoptosis can proceed without Mcl-1 degradation. *J Cell Biol* 180, 341-55.
10. Day, C. L., Smits, C., Fan, F. C., Lee, E. F., Fairlie, W. D. & Hinds, M. G. (2008). Structure of the BH3 domains from the p53-inducible BH3-only proteins Noxa and Puma in complex with Mcl-1. *J Mol Biol* 380, 958-71.
11. Sattler, M., Liang, H., Nettesheim, D., Meadows, R. P., Harlan, J. E., Eberstadt, M., Yoon, H. S., Shuker, S. B., Chang, B. S., Minn, A. J., Thompson, C. B. & Fesik, S. W. (1997). Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. *Science* 275, 983-6.
12. Frank, R. (1992). Spot-synthesis - an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron* 48, 9217-9232.

13. Liu, X., Dai, S., Zhu, Y., Marrack, P. & Kappler, J. W. (2003). The structure of a Bcl-xL/Bim fragment complex: implications for Bim function. *Immunity* 19, 341-52.
14. Fire, E., Gulla, S. V., Grant, R. A. & Keating, A. E. (2010). Mcl-1-Bim complexes accommodate surprising point mutations via minor structural changes. Protein Sci 19, 507-19
15. Day, C. L., Chen, L., Richardson, S. J., Harrison, P. J., Huang, D. C. & Hinds, M. G. (2005). Solution structure of prosurvival Mcl-1 and characterization of its binding by proapoptotic BH3-only ligands. *J Biol Chem* 280, 4738-44.

Chapter 5

Conclusions

In this thesis I showed applications utilizing both computational and experimental methods for the design of protein-protein interaction specificity. Below I will first review briefly what we learned from these applications, and suggest possible future improvements. To conclude, I will suggest a general framework to combine different approaches in these studies, especially those presented in Chapter 2 and Chapter 3, to enhance our future design capabilities.

Summary of design applications in this thesis

In Chapter 2 I used a scoring function specific for bZIP coiled coil to computationally design a peptide inhibitor targeting the BZLF1 protein. One major question was whether the scoring function was robust enough given the non-canonical sequence and structural features of BZLF1. Through the success of the design and other mutational studies, we showed that at least some of the key specificity determinants learned from more typical human bZIP coiled coils could be applied for this system as well, probably by making the designed model more canonical in the re-modeled region. The demonstration of such "modularity" is encouraging for future studies that aim to perform protein design by applying experimental knowledge obtained from related proteins.

In Chapter 3 I used structural modeling to guide the design of a library to be screened for anti-apoptotic Bcl-2 proteins with novel interaction specificities. I suggested an approach to broadly include predicted non-disruptive residues in addition to specific ones, providing a "safety net" for important residues missed by the difficult specificity predictions. To fully utilize the power of library screening, I developed an optimization framework to maximize inclusion of

the predicted non-disruptive residues under a constraint of library size. Analyzing specific sequences obtained from the screening revealed that key important residues were missed by our specificity predictions but captured by our non-disruptive residue predictions. We also showed that contributions among different residues can be highly non-additive. As higher order interactions among different residues can be difficult to predict, we argue that this demonstrates the importance of allowing combinatorial possibilities to be explored among residues at different designed positions.

Reflecting on results obtained from Chapter 3, there appear to be several aspects of the library design approach outlined above that can be further improved upon. In using structural models to guide library design, we relied only on the Rosetta modeling suite. Although no current modeling protocol has been shown to be vastly superior over others in terms of specificity predictions, examining prediction results from multiple models could potentially remove biases generated by a single model. For example, specificity features predicted by multiple models could be considered for inclusion with higher priority over ones predicted by only one model. Or alternatively, one could enforce the inclusion of any specificity features predicted by any model. However, I observed during the library optimization phase that the constraint to include all predicted specificity residues could lead to usage of inefficient degenerate codons, significantly compromising the inclusion of predicted non-disruptive residues. Relaxing this constraint can potentially lead to designed libraries with better quality. This can be done by allowing the omission of one or more predicted specificity residues, or by the approach presented in Chapter 4 in which predicted specific residues are given more weights in the optimization, but not necessarily always included. Finally, in this study we pre-determined the number of designed positions before library optimization. We also enforced a hard constraint on

library size. It would be desirable, in future applications, to carry out a more rigorous exploration of how including different numbers of positions and relaxing the library size constraint would affect the quality of the library, before choosing one library for screening.

In Chapter 4 I developed a PSSM model from SPOT array data and used it to analyze and design specificity of BH3 peptides binding different anti-apoptotic Bcl-2 proteins. We showed that despite the many deficiencies, the simplified model could nonetheless provide interesting insights into sequence determinants of specificity. In the future, it will be interesting to investigate whether such models can be further improved by combining terms derived from structural modeling or from other experimental data such as library screening.

New design framework

From a more application oriented viewpoint, the bZIP coiled-coil scoring function developed by Grigoryan et al.[1] and used in Chapter 2 suggested an intriguing solution to bypass the current limitations of more general scoring functions. The function contains components derived from structural models as well as experimental information. It is specific for use with coiled coils, particularly bZIP coiled coils, and cannot be applied to other types of proteins. On the other hand, the restrictive nature also allows interaction specificity to be captured with higher accuracy under a simpler formulation, which can be incredibly useful for design. One challenge for this approach is that the protein family of design interest might not have natural representatives that span a diverse sequence and interaction profile space. This makes it difficult to extract information useful for design from interaction data obtained from these proteins, using methods such as machine learning[2,3,4]. Taking the design application in Chapter 3 as an example, all natural human anti-apoptotic Bcl-2 proteins bind the BH3 peptide Bim strongly. It is therefore less likely that interaction data among natural anti-apoptotic Bcl-2 proteins and BH3 peptides will encode

information about possible sequence determinants specific against Bim. In this regard, exploring "unnatural" sequence and interaction space becomes necessary. One solution to address both issues is to screen large random libraries for interaction properties not observed in nature. However, due to the combinatorial nature of sequence space, such practice might be more suitable for protein interfaces involving a smaller number of residues (e.g. a small protein interacting with a short peptide)[5,6].

In Chapter 3, I described efforts to computationally design a library of sequences, reasoning that structural models can enhance the probability of identifying sequences with the desired interaction specificities in a library screen. Although not explicitly demonstrated in this thesis, such a guided library screening approach would be more efficient in creating a meaningful interaction data set compared to screening a random library as described in the previous paragraph. In the study in Chapter 3, we explicitly screened for and identified sequence determinants that favored binding Bad over Bim. Interestingly, one of the specific sequences actually displayed global specificity against other BH3 peptides, even though such specificity was neither designed nor screened for. As discussed in Chapter 3, this could happen simply by chance. However, it is also intriguing to think that a library design strategy guided by rather simple principles (enriching sequences predicted to bind Bad and including predicted specificity elements against Bim) can help enrich sequences with other types of specificity elements as well. Obviously, this represented only a minor step toward comprehensively mapping sequence and specificity space. Nonetheless, we can extend this approach to identify sequences specific for BH3 peptides other than Bad. The goal of a complete understanding of the relation between sequence and specificity is difficult to attain, but garnering information on various sequence elements that favor or disfavor binding to each BH3 peptide can be within reach. This knowledge,

combined with structural models that fill in the missing terms not learned from experiments, can be sufficient for the task of designing individual sequences with other novel interaction specificity patterns, without the need for repeated library screening efforts.

In summary, the approach suggested above consists of the following steps: (1) Use structural modeling to make guided libraries, (2) Screen the libraries for sequences with interaction specificities that one wants to learn, (3) Try to extract information from the screening results and combine that with structural models to build better scoring functions, and (4) Use the better model to design other desired interaction specificities. My thesis demonstrated the first two points. More computational and experimental studies will be required to test the feasibility of the latter two.

The above represents a general outline for the proposed design framework. It is likely that one will meet different problem-specific challenges when actually going through the whole process. One potential challenge arises if the protein interaction interface of interest is highly flexible. Scoring functions derived from experiments usually assume an invariant set of important contacts between certain positions at the interface (determined from available structures of the complexes of interest), and the strength of an interaction can be estimated as the sum of weights for these contacts. Such formulation makes it convenient to train or parametrize weights for these contacts. However, this assumption is less valid if major conformational changes at the interface are involved. Contacts can occur between significantly different interfacial positions instead to generate different binding models. Such possibility should be considered when available data is in conflict with a scoring model that dictates only one rigid protein interface. Learning potential conformational changes from interaction data alone (point 3 above) is a great challenge, and knowledge of different binding models can be helpful. If no

existing structural data is available, the same tools to explore backbone flexibility as describe in Chapter 1, albeit no substitute for experimental structural data, can be used to obtain insight. Obtaining structural data for complexes suspected to adopt different binding models can also be extremely valuable. More sophisticated machine learning algorithms[4] can then be applied to learn from interaction data while incorporating such information.

References

1. Grigoryan, G., Reinke, A. W. & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859-64.
2. Fong, J. H., Keating, A. E. & Singh, M. (2004). Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol* **5**, R11.
3. Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A. & MacBeath, G. (2008). Predicting PDZ domain-peptide interactions from primary sequences. *Nat Biotechnol* **26**, 1041-5.
4. Gfeller, D., Butty, F., Wierzbicka, M., Verschueren, E., Vanhee, P., Huang, H., Ernst, A., Dar, N., Stagljar, I., Serrano, L., Sidhu, S. S., Bader, G. D. & Kim, P. M. (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol* **7**, 484
5. Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D. & Sidhu, S. S. (2008). A specificity map for the PDZ domain family. *PLoS Biol* **6**, e239.
6. Tonikian, R., Xin, X., Toret, C. P., Gfeller, D., Landgraf, C., Panni, S., Paoluzi, S., Castagnoli, L., Currell, B., Seshagiri, S., Yu, H., Winsor, B., Vidal, M., Gerstein, M. B., Bader, G. D., Volkmer, R., Cesareni, G., Drubin, D. G., Kim, P. M., Sidhu, S. S. & Boone, C. (2009). Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* **7**, e1000218.